# Artificial Intelligence and Social Scoring Systems: Between Dystopia and Reality

## *Inteligencia artificial y sistemas de puntuación social: entre la distopía y la realidad*

**Ginevra Cerrina Feroni**
Catedrática de Derecho Constitucional Italiano y Comparado
Universidad de Florencia
Vicepresidenta del Garante per la protezione dei dati personali

**Resumen:**

Este trabajo examina los sistemas de puntuación social y de reputación algorítmica como una de las manifestaciones más intensas de la gobernanza automatizada contemporánea. A partir del caso paradigmático de la República Popular China y de su Sistema de Crédito Social, el artículo analiza la lógica de clasificación y vigilancia que subyace a estos mecanismos, así como sus efectos sobre las oportunidades, libertades y condiciones de vida de las personas. El estudio aborda también el caso neerlandés del *Toeslagenaffaire*, en el que el uso de algoritmos en la detección de fraude produjo efectos discriminatorios y decisiones gravemente lesivas, y la experiencia italiana en materia de rating reputacional, especialmente a la luz de la jurisprudencia y de las decisiones de la Autoridad de Protección de Datos. Sobre esta base, el trabajo sostiene que la expansión de la toma de decisiones automatizada exige reforzar las garantías de transparencia, control humano, proporcionalidad y protección de la dignidad de la persona. La reflexión final propone desplazar la atención desde la algocracia hacia una algorética orientada por los derechos fundamentales.

**Palabras clave:**

Inteligencia artificial; puntuación social; crédito social; rating reputacional; protección de datos.

***Abstract:***

*This article examines social scoring systems and algorithmic reputational rating as one of the most far-reaching manifestations of contemporary automated governance. Taking the People's Republic of China and its Social Credit System as a paradigmatic case, the paper analyses the logic of classification and surveillance underlying these mechanisms, as well as their effects on individuals' opportunities, freedoms, and living conditions. It also addresses the Dutch Toeslagenaffaire, where the use of algorithms in fraud detection generated discriminatory outcomes and seriously harmful decisions, and the Italian experience with reputational rating, especially in light of case law and decisions of the Data Protection Authority. On that basis, the article argues that the expansion of automated decision-making requires stronger guarantees of transparency, human oversight, proportionality, and protection of human dignity. The final section advocates a shift from algocracy to an algoretics grounded in fundamental rights.*

***Keywords:***

*Artificial intelligence; social scoring; social credit; reputational rating; data protection.*

**Sumario:**

**1.** Una distopía que ya es realidad. **2.** La República Popular China: el Sistema de Crédito Social. **2.1.** Planificación e implementación. **2.2.** Factores para evaluar el comportamiento de los ciudadanos. **2.3.** Efectos del Sistema de Crédito Social. **3.** Los Países Bajos: el *Toeslagenaffaire*. **3.1.** Los hechos. **3.2.** Aspectos jurídicos relevantes. **4.** La experiencia italiana. La calificación reputacional en la jurisprudencia de la Autoridad de Protección de Datos. **4.1.** El caso *Mevaluate*. **4.2.** El contexto escolar. **5.** Reflexiones finales. De la algocracia a la algorética.

***Summary:***

***1.*** *A dystopia that is already reality.* ***2.*** *The People's Republic of China: the Social Credit System.* ***2.1.*** *Planning and implementation.* ***2.2.*** *Factors for evaluating citizens' behavior.* ***2.3.*** *Effects of the Social Credit System.* ***3.*** *The Netherlands: the Toeslagenaffaire.* ***3.1.*** *The facts. 3.2. Relevant legal aspects.* ***4.*** *The Italian experience. The reputational rating in the case law of the Data Protection Authority.* ***4.1.*** *The Mevaluate case.* ***4.2.*** *The school context.* ***5.*** *Concluding reflections. From algocracy to algoretics.*

## 1. A dystopia that is already reality

Let us imagine if every one of our actions, interactions, movements were reduced to an assessment on a five-point scale. A higher score opens the doors to fabulous opportunities and special advantages, while a low score can, essentially, isolate us from the rest of society. Thus, for example, by reason of how we answered the waiter at the café where we ordered a coffee, it could be forbidden for us to get on public transport, or depending on the evaluations of our previous neighborhood, the possibility could be precluded of moving to another district populated by subjects whose score is higher than a certain threshold. The subject of one of the most watched episodes of the science-fiction series "Black Mirror", this, however simplified, is how a system of social score or social credit works, that is, of social scoring or reputational rating.

But what still sounds like the dystopian plot of a science-fiction narrative is a reality that already surrounds us and directly concerns the citizens of some regions of the world.

One of the best-known and most evident social credit systems is the one introduced by the Chinese Government, although China is not the only legal system that is adopting solutions for social monitoring. In January 2022, the government of the United Kingdom revealed that it would use real-time facial recognition software on the streets of London to find suspects wanted by the police, and similar systems are already used by law-enforcement agencies and intelligence agencies around the world. In addition to public authorities, other entities, such as some insurance companies, have announced or already implemented systems that would allow their operators to effectively collect information about people's behavior to be used in important decisions. For example, companies specializing in life insurance, in New York, are authorized to make decisions about their clients using information found on social networks. After all, the automobile insurance sector already uses black boxes installed on cars as part of the contract, capable of calculating the malus on the basis of the driving style that they perceive.

There are programs that allow owners of restaurant or entertainment businesses to create lists of people not authorized to enter because of previous improper behavior. And all sharing-economy services, such as Airbnb, Uber or home delivery services, have a kind of scoring system that evaluates those who participate in the service both as workers and as consumers. With the Covid-19 health emergency, the issue of monitoring citizens imposed itself in an even more pervasive manner, concentrating upon itself the radical cultural divergences, before political and legal ones, regarding the possibility that regulatory authorities around the world could or could not hypothesize solutions to trace infected or unvaccinated citizens and, on the basis of their health data, allow or prohibit for them access, opportunities and even the very exercise and enjoyment of constitutionally guaranteed rights and freedoms.

Hannah Arendt opens Part Three of her fundamental work on the origins of totalitarianism with a quotation from David Rousset: "ordinary men do not know that everything is possible"[1]. The real risk, therefore, is not that of failing to notice that certain scenarios that we did not even know we could imagine are today reality and thus failing to protect ourselves from them, but precisely of failing to notice them because we are getting used to them. To address the topic it is worth dwelling, albeit in broad strokes, on the to-date most developed experience of a social credit system, namely the case of the People's Republic of China. We will then go on to recall a social credit affair that concerned the Netherlands and was the subject of a broad public debate, and then analyze the Italian case, also in the light of the most significant case law of the Data Protection Authority.

[1] Hannah Arendt, *The Origins of Totalitarianism* (New York: Schocken Books, 1951), 421.

## 2. The People's Republic of China: the Social Credit System

The Social Credit System (SCS) is an initiative developed by the People's Republic of China since the early two-thousands with the aim of developing a national system to classify the reputation of its citizens. Through the SCS each citizen is assigned a score representing his "social credit," on the basis of information held by the Government, such as those regarding his economic and social condition, his preferences and his habits, thus operating essentially as a mass-surveillance mechanism entrusted to AI technologies for the analysis of big data[2]. In addition to individual citizens, as will be seen, the SCS has also been extended to the activities of all enterprises operating in the Chinese market.

### 2.1. Planning and implementation

The Social Credit System is placed within the Chinese so-called "top-level design" approach and is coordinated by the "Central Leading Group for Comprehensively Deepening Re-forms". According to the "Planning for the Establishment of the Social Credit System (2014-2020)" published by the State Council, integrated with the recent "Social Credit Action Plan 2024-2025"[3], the Social Credit System focuses on four areas: "honesty in government af-fairs," "commercial integrity," "social integrity," and "judicial credibility".

The objective of the initiative is twofold: on the one hand there is the individual plan that per-tains to the ethical sphere of the actions of the individual: in this dimension the SCS is aimed at identifying and punishing deviance and at incentivizing sincerity, honesty, and integrity of individual Chinese citizens[4]. On the other hand - and perhaps this is the most neglected aspect of the phenomenon - the intent is expressly declared to create prodromal conditions for the perfection and regulation of the economy of a socialist market[5]. The strengthening of the State's dirigiste capacity and its ability to steer the behavior and actions of citizens are in fact aimed not only at needs of social control, but also at the need for uniformity of the Chinese market. From this latter point of view the SCS project pursues two fundamental economic objectives: on a small scale, to eliminate market pathologies, such as fraud and counterfeiting; on a large scale, to solve the problem of financial reliability within and by the Chinese market and, in this way, to contribute to realizing the union between social harmony and the harmo-nious economic development of the Country[6].

The Social Credit System, for now, appears confined to mainland China, and not also to Hong Kong and Macao. Moreover, the plans for the enlargement and development of the system do not speak of a distinction between Chinese enterprises and foreign enterprises op-erating in the Chinese market, implying the possibility that foreign enterprises operating in China are also subjected to classification within the system.

The system has already been involved in quite a few controversies, particularly with regard to the way in which it will be applied to people and to enterprises, and yet neither the cases won before the courts nor the errors made by the algorithm and recognized as such by the authori-ties themselves have led to a modification of the process of functioning of the same[7].

As regards citizens, examples are already detectable of punishments caused by the violation of social protocols: the system has already denied nine million people with "low scores" the right to

[2] Among the most recent works that have attempted to systematically describe and define the topic: C. Liu and A. Rona Tas, "Trusting by Numbers: An Analysis of a Chinese Social Credit System Governance Infrastructure," *Critical Sociology* 51, no. 6 (September 2025): 1247–1265; C. Loefflad, M. Chen, and H. Grossklags, "Reputational Discrimination and Fairness in China's Social Credit System," *Digital Government* 5, no. 4 (December 2024): 1–27; see also X. Dai, "Toward a Reputation State: A Comprehensive View of China's Social Credit System Project," in *Social Credit Rating*, ed. O. Everling (Wiesbaden: Springer, 2020), 139–165; B. Ahl, L. Catá Backer, and Y. Chen, "Law and Social Credit in China," *China Review* 24, no. 3 (2024): 1–15; S. Hofmann, *Social Credit: Technology-Enhanced Authoritarian Control with Global Consequences*, International Cyber Policy Center, Policy Brief Report no. 6 (2018).

[3] National Development and Reform Commission, *Social Credit Action Plan 2024–2025* (June 2024), testo in inglese su https://www.chinalawtranslate.com/en/2024-2025social-credit-plan/

[4] Not by chance, R. Creemers, *China's Social Credit System: An Evolving Practice of Control*, SSRN, 2018, 2, notes that in Mandarin the word "credit" (xinyong) is semantically identical to the term "integrity".

[5] B. Ahl, L. Catá Backer, Y. Chen, "Law and Social Credit in China", cited work, esp. pp. 7 ff.

[6] L. Yu-Hsin Lin, C. Milhaupt, *China's Corporate Social Credit System: The Dawn of Surveillance State Capitalism?*, The China Quarterly, 256, 2023, 835–853, esp. 836; see also D. Mac Sithing, M. Siems, *The Chinese Social Credit System: a Model for other Countries*, Modern Law Review, 82(6), 2019, 1034-1071, esp. pp. 1048 ff.

[7] M. van Blomberg, *The Social Credit System in China's Rule of Law*, Mapping China Journal, no. 2, 2018, pp. 78-162, esp. p. 100.

buy domestic flights. The system is also used to exclude the possibility for parents to enroll their children in certain schools, to prohibit those who have a low score from renting hotel rooms, from using credit cards, and to insert certain individuals on a blacklist with the aim of precluding for them the obtaining of work[8]. The system has also been used to evaluate people's internet browsing habits (too much time spent playing video games reduces the score for example), purchasing habits and a variety of personal and absolutely harmless ac-tions that have no impact on the community[9].

The "Notice of the State Council regarding the issuance of the 'Planning for the Establish-ment of the Social Credit System (2014-2020)'" was issued by the Chinese State Council on June 14, 2014. In 2015 a license was granted to eight companies so that they could begin to develop prototypes of credit systems. The selected enterprises are Ant Financial of the Aliba-ba Group, the software developer Tencent and six others. The credit systems created by these enterprises used databases such as Sesame Credit at the beginning of the experimentation phase[10].

Until now, there exists no comprehensive social credit system extended to the whole nation, but multiple experimental projects have been implemented that are testing the system on a local scale and in some sectors of the economy. A program of this type was implemented in Shanghai by means of the Honest Shanghai app, which uses facial recognition software to search within government archives and classify users in this way. The scores can also be based on information extrapolated from the online search habits of Chinese citizens.

The infrastructure of the Social Credit System was completed in 2020 and strengthened in recent years, but the restrictions on citizens and enterprises with "low scores" began to take effect starting already in 2018.

With regard specifically to enterprises, instead, as already said, the Social Credit System was designed as a mechanism of market regulation. The objective is to establish a structure of self-imposed regulation, powered by big data, within which enterprises control each other. The basic idea is that with a functional credit system active, companies will endeavor to comply with Government policies and provisions in order to avoid their score going down. As it was conceived, enterprises with high scores will have advantages such as better conditions on loans, lower taxes and more investment opportunities. Enterprises with low scores will re-ceive disadvantageous conditions for new loans, higher taxes, restrictions on investments and fewer opportunities to participate in projects financed by the public sector. The Government's plans also provide for real-time monitoring of the activities of enterprises. In that case, in-fringements by an enterprise could take shape in an instantaneous lowering of the score[11].

## 2.2. Factors for evaluating citizens' behavior

The Chinese social credit system is based on three variables:

1. A first factor, not dissimilar indeed from that taken into consideration in "private" social credit systems (e.g., Schufa in Germany, or FICO in the USA) is that linked to the ability of each citizen to regularly fulfill his own financial obligations. A person who pays his bills punctually can be considered reliable and solvent and for this reason will receive a higher score.

2. A second factor concerns the ability of each individual to fulfill his own duties as defined in the contracts entered into. For example, an individual with a substantial level of savings will receive a higher score than a person who struggles to make it to the end of the month.

3. A third factor, the one that presents the greatest criticalities, concerns social behavior extensively considered: purchasing orientations on online portals such as Alibaba, the trends of offline purchases paid with Alipay, social interactions on WeChat, the online history of Baidu, the areas of frequentation (identified through geolocation of the cell phone), that is in general those elements that can give an all-encompassing idea of the personality of each citizen. For example, a user who buys diapers is very likely a parent and, therefore, also a more responsible citizen. On the contrary, a citizen who spends large sums on online games is considered lazier and less productive for society. It is easy to understand how such results can be obtained only by an algorithmic evaluation system

[8] Among the many who reported these examples, D. Mac Sithing, M. Siems, *The Chinese Social Credit System: a Model for other Countries*, cited; and B. Ahl, L. Catá Backer, Y. Chen, *Law and Social Credit in China*, cited.

[9] A. Fenwick Elliott, *China is banning people with bad 'social credit' from using planes and trains*, The Telegraph, 19 March 2018.

[10] M. van Blomberg, *The Social Credit System in China's Rule of Law*, cited, p. 93 (Sesame Credit description).

[11] See B. Ahl, L. Catá Backer, Y. Chen, *Law and Social Credit in China*, cited, esp. pp. 7 ff.

whose source code is highly influenced by social values of collective interest[12].

### 2.3. Effects of the Social Credit System

The Chinese system does not limit itself to analyzing the behavior of the individual[13]. It is in fact provided that each individual's score is also influenced by the behavior of the people closest to him. For example, a subject who violates one or more laws of the Chinese Communist Party will also negatively influence partners and persons connected to him[14].

The effects of the Social Credit System can concern a series of activities.

a) Flight bans: currently, the flight ban for persons considered unreliable is already common practice in China[15].

b) Exclusion from private schools: if the parents' score were to be below a certain threshold, their children would be excluded from the best schools in the region.

c) Slowing of the internet connection: citizens considered unreliable could see their internet connection slowed down or even be completely excluded from access to specific websites.

d) Exclusion from high-prestige jobs: the score of the Chinese SCS could in the future be part of a person's curriculum vitae and thus provide essential information to potential employers. In this way for people with a low score, it would almost surely be [impossible] to reach the most sought-after and qualified work positions.

e) Registration on a public blacklist: a prototype of a blacklist currently exists in China and all those who are registered on such a list could expect one or more of the penalties mentioned previously.

Another crucial step in the process of constructing the SCS is also represented by the parallel reward system that is envisaged for citizens aimed at incentivizing the latter to behave according to the rules in order to be able to enjoy facilitations with reference to the same type of activities that are forbidden to others. The "virtuous circle" that follows consists above all in the advantage that the Chinese State is not required to provide any direct intervention either of control or corrective on this type of activities[16]. Among the reward measures, one can recall facilitated access to financing (in the Sesame Credit pilot project, those who reach a score of 600 can request a loan up to an amount of 5,000 Yuan) and to rentals or leases (including exemption from the obligation of a security deposit), facilitation of travel (depending on the score, fewer documents are required to justify the request for travel visas) and the increase in social status that follows its official recognition (one's own score can be used as a status symbol on social platforms and dating apps)[17].

## 3. The Netherlands: the *Toeslagenaffaire*

Although the Chinese case of profiling on a reputational basis certainly remains the most extreme, nonetheless it constitutes only one part of a broader phenomenon. One cannot fail to note, in fact, that, favored by the growing power of the algorithm, the use of social scoring on the public-law level has found its first applications on a global scale and, indeed, in Europe. The most striking case, which in January 2021 led to the resignation of the then Prime Minister Mark Rutte, concerns the so-called *Toeslagenaffaire*, by reason of which it was brought to light that the Dutch tax administration and, in particular, its *Toeslagen* unit (the welfare department) had unlawfully requested the retroactive repayment of subsidy allowances received by about 35,000 parents. It was precisely the use of the unlimited automation of machine-learning algorithms that played a significant role in the scandal which, moreover, caused the Dutch public finances damage of about 500 million euros, thus revealing the uneconomical nature of the entire operation.

[12] Hence the co-ordinate value that Chinese civilization attributes to social norms and to law. See Z. Zuo, *Governance by Algorithm: China's Social Credit System*, University of Cambridge Working Paper, 16 June 2020.

[13] J. Chin and L. Lin, *Surveillance State*, St. Martin's Press, 2022, pp. 219 ff.

[14] *China's Chilling Social Credit' Blacklist*, Human Rights Watch, 12 December 2017.

[15] *China Releases Investigative Journalist After Almost Year in Jail*, Newsweek, 3 August 2014.

[16] M. van Blomberg, *The Social Credit System in China's Rule of Law*, cited, esp. p. 93 ff.

[17] As reported by S. Johnson, How China's 'social credit score' will punish and reward citizens, Big Think, 26 April 2018, the Baihe website, for example, already allows its users to publish their own score.

## 3.1. The facts

Towards the end of 2011 the Dutch press revealed a series of cases of fraud against social security concerning payments made by the tax administration in favor of people who resided in other EU Member States[18]. This series of cases, christened by the press *Bulgarenfraude* (Bulgarian fraud)[19], thus represented the incentive to reform the welfare system, in order to increase the efficiency of the process of identifying fraud and recovering undue emoluments. This streamlining was pursued through various means including the implementation of machine-learning algorithms to automate the identification of culpable errors or fraud in the processing of social-security forms submitted by parents. In practice, however, the same sanctions, namely the suspension of allowances and the request for restitution of what had been unduly received, were imposed not only for intentional ideologically false crimes, but also for material errors and simple formal oversights, such as, for example, the erroneous completion of a box or the omission of a signature. This was due to the fact that Art. 26 of the Income Schemes Act (AWIR) did not contemplate the obligation for the administration to apply the principle of proportionality[20].

The use of this AI algorithm for "zero-tolerance" risk detection therefore obliged the full repayment of the debt with no possibility of its temporal deferral and often with the addition of the payment of a surcharge, even for individuals to whom only material errors in completing the application forms for the welfare contribution could be imputed.

The welfare department, moreover, had used a machine-learning algorithm also in the risk-assessment process for the selection of beneficiaries of childcare allowances to be subjected to specific checks. As is typical of machine-learning algorithms, the system inferred the existence of risk factors on the basis of the analysis of historical data, that is, of previous known cases of fraud. This is, as has been known for some time, a method that is not free from risks of producing biases and discriminations, which for this purpose requires a twofold protocol of safeguards: first of all the data processed both during and after the automatic-learning process must accurately represent the target population, that is, the recipient of the welfare measure; secondly the risk factors (the so-called "weights" in the terminology of machine learning) must be as free as possible from any kind of influence that concerns the conduct, generally considered, held by minorities or by individuals of lower socio-economic status. It is no coincidence that, when it is said that algorithms (and also predictive algorithms) are based on decisions already taken, this means that it is not the algorithm that contains the bias, since what is discriminatory is the historical sequence of decisions that lies upstream. For these reasons, circumventing such distortion requires vigilance, constant monitoring, spot checks and also a certain degree of skepticism with regard to the conclusions of the algorithm[21].

In the *Toeslagenaffaire* the Dutch Data Protection Authority (AP) and the National Audit Service (ADR) showed how the attitude held by the tax administration was in reality of a completely opposite sign. The *Toeslagen* unit would not have exercised prudent oversight nor would it have verified whether the algorithm was free of bias, indirectly inducing the algorithm to return distorted results. Proof of this is the so-called "twin test"[22] carried out by the AP and from which it emerged that, all other conditions being equal, the algorithm flagged a higher risk of fraud on the part of individuals who were not of Dutch nationality compared to Dutch citizens who presented similar characteristics[23].

[18] According to calculations by the Dutch Data Protection Authority (Autoriteit Persoonsgegevens), "Werkwijze Belastingdienst in strijd met de wet en discriminerend," July 17, 2020, the events led to net revenue losses of about €10 million for the Belastingdienst, while the Dutch government had to step in by creating a €500 million fund to compensate the victims.

[19] The label was linked to the fact that some of these fraud cases involved families of Bulgarian nationality: "*Uitspraak Bulgarenfraude: hoe zat het ook alweer?*", on RTL Nieuws, 19 May 2015.

[20] See also Venice Commission, Council of Europe, Opinion no. 1030/2021, p. 5. The legitimacy of that law had been upheld on several occasions by the Dutch Supreme Administrative Court, which deemed lawful the failure to apply the principle of proportionality to the seriousness of the offence.

[21] Parlementaire ondervragingscommissie Kinderopvangtoeslag, *Tweede Kamer der Staten-Generaal* 11 July 2020.

[22] A "twin test" is a simple procedure in which two identical fictitious profiles are created, differing only by one characteristic; in this case, the characteristic was "Dutch/non-Dutch."

[23] Thus the Autoriteit Persoonsgegevens, "*Belastingdienst beloofde ambtenaren niet te straffen om toeslagenaffaire*", 20 May 2019, according to which, in substance, the algorithm discriminated against foreign residents. In its report, the AP describes a particularly troubling case in which, based on alerts concerning a specific childcare institution where about 100 parents of Ghanaian origin were suspected of fraud, the *Toeslagen* unit arbitrarily decided to investigate all 6,047 parents of Ghanaian origin living in the Netherlands. Under such conditions - where supposedly "objective," "data-driven" decisions are in fact arbitrary, biased, and discriminatory - it is hard to deny that the machine adopted a discriminatory decision, much as human tax officials might have done. Moreover, the case of the Ghanaian parents is not the only instance in which foreigners were subjected to an arbitrary decision by the leadership of the *Belastingdienst*.

Adding up little by little, these arbitrary manual targetings of foreigners in the machine-learning system, which learns and constantly updates the risk factors on the basis of the historical data entered by the Tax Administration officials, induced a veritable distortion in the algorithm, which heuristically concluded that welfare beneficiaries of foreign origin and citizenship were more inclined to fraud. Even more problematic is the fact that welfare beneficiaries, considered potential fraudsters, suffered the interruption of the benefit without even any *ex post* assessment by human operators.

### 3.2. Relevant legal aspects

The first lesson to be drawn from the Toeslagenaffaire is that one cannot automate a process that has original flaws, since automation will only amplify its scope. Machine-learning algorithms were used in the Toeslagenaffaire to reduce the serious backlog in the process of recovering the childcare allowance. However, since the recovery process was in the first place gravely flawed - particularly due to the lack of application of the principle of proportionality in the case of material errors by welfare beneficiaries - the algorithms determined a result completely opposite to what was envisaged: that is, they further increased the backlog at the *Belastingdienst* and the number of appeals brought by welfare beneficiaries against the Toeslagen unit[24].

The second conclusion is that States should use machine learning and automation by arranging ex ante the legal and technological conditions so that these algorithms can be used lawfully and ethically. Well before the Toeslagenaffaire, it had been repeatedly pointed out how automatic learning was inclined to create serious risks of biases and discriminations, as has repeatedly emerged over the last years, for example with reference to the right to a fair trial, to good administration and to other principles derivable from constitutional or interposed norms[25]. Starting from this, the adoption of machine-learning algorithms, in the absence of strengthened safeguards and the protection of these rights, is inevitably the harbinger of perverse effects.

Which leads to the third lesson to be gleaned from the Dutch case: the necessary control by the human operator. The systematic adoption of machine-learning tools roughly coincides with the global financial crisis of 2008-2011, in which austerity measures led to a reduction in the workforce of tax administrations throughout the European Union. This should in itself be a cause for concern, since it is likely that a reduction in staff affects the number of officials in charge of taxpayers' ex post complaints, and therefore their review capacity. By doing so, we are not creating the conditions for an anthropocentric AI, but rather a decision-making system centered on uncontrolled artificial intelligence, completely antithetical to the objectives that one intends to promote with algorithmic-governance projects of a Community stamp.

The reported affair is, therefore, of great interest in that it is easy to imagine that given the prerequisites that determined it, it is not an isolated phenomenon, but that, on the contrary, it may be repeated in other EU Member States.

## 4. The Italian experience. The reputational rating in the case law of the Data Protection Authority

The proliferation of applications or private databases that indiscriminately collect and store personal data on the net in order to make decisions also with negative effects and consequences on the private, social and economic life of individuals has also reached our country. These projects aggregate information coming from different sources, not always of an objective nature - e.g. also posts from social profiles - and extend the evaluation model, now consolidated and applied to accommodation or restaurant facilities, directly also to the evaluation of people.

These are systems that aim, in substance, to make measurable the reputation of the registered subjects, raising more than one concern in terms of personal-data protection, notwithstanding that the declared purposes are almost always justified by the pursuit of significant objectives such as, for example, the fight against corruption, against false identities, against fraud, etc. But the very idea of entrusting the "review" of a person to an algorithm risk truly opening a very dangerous drift. Hence a reflection is required right from the start that goes beyond the technicist application of the

[24] M. Fenger, R. Simonse, *The implosion of the Dutch surveillance welfare state, in Social Policy & Administration*, vol. 58, Issue 2, 2024, pp. 264-276.

[25] [27] C. Castro, *What's Wrong with Machine Bias*, Ergo, 5(15), 2019–2020; J. Angwin et al., Machine Bias, ProPublica, 23 May 2016; E. Stradella, *Stereotipi e discriminazioni: dall'intelligenza umana all'intelligenza artificiale*, in *Consulta Online*, 20 march 2020, online su www.giurcost.org, in part. pp. 9 ff.; A. Gatti, *L'algoritmo tra volontà e rappresentazione*, in *DPCE online*, n. 3, 2020, pp. 3457-3461.

GDPR rules (and in particular of Art. 22, on automated decisions including profiling and subsequent ones on security), to dwell rather on the general principles enshrined by it.

Reputation is a complex element and proper to an entire personality, corresponding to several aspects that concern multiple personal data that are also significant and sensitive; therefore, as such it is protected by the GDPR and by the Privacy Code and is guaranteed by an independent Authority. The prejudice to reputation is expressly listed among possible damages consequent to a personal-data violation of the data subject, pursuant to Recital 85 of the GDPR. An emblematic case can occur precisely in the hypothesis of publication of personal data processed for profiling purposes.

It is evident that the reputation of individuals that one would like to quantify in order to assign reliability scores is closely linked to the person considered in his social projection and is intimately connected with dignity, a keystone of the discipline on personal-data protection, which looks precisely at the impact of inconsiderate processing, especially if automated and of profiling, on the rights and freedoms of individuals, paying attention so that the data subject can always exercise his own control over his data processed by others[26].

Many doubts concern the actual quality of the data collected - destined precisely to be processed by specific algorithms - or the genuineness of any reviews entered by third parties that can easily lend themselves to distorted, defamatory or harmful uses. Judgments that risk seriously compromising the right to the integrity of personal identity and whose prejudicial effects are irreparably amplified by users' free access through search engines. And crystallized in the hypermnesia of the net for an indefinite time.

Further evaluations are imposed both by the concrete methods of managing such rating systems and by the discretion of the measurement criteria identified by the companies that intend to offer such services on the market. It is no secret that in the past managers, in some cases, even knowingly altered the reference market by inserting fake reviews. These are projects and practices that, in the face of the various critical issues raised, have a considerable impact on the personal sphere of users, on their right to informational self-determination. This dimension, whose centrality has been reaffirmed by the case law of the Court of Justice of the European Union, which since 2014 has recognized the exercise of a new right to digital oblivion, is therefore crucial[27].

A human being is not a company name, a commercial asset, a service and cannot and must not become one[28]. The Court of Luxembourg has established that, in the event of conflict, the mere economic interest succumbs, when certain conditions occur, in relation to the category of fundamental rights that are an expression of the principles of dignity contained in the Charter of Rights[29].

### 4.1. The Mevaluate case

With measure no. 488/2016, the Privacy Authority judged as unlawful the platform of the "Mevaluate Onlus" Association which collects numerous categories of personal data (pertaining to the following spheres: criminal, tax, civil, work and civic engagement, studies and training) for the calculation of a reputational rating of the applicant (natural or legal person), in order to "make socio-economic relations more efficient, transparent and secure". Such information is collected on the basis of the consent of the data subjects and is provided by sources other than the data subjects, possibly supplemented by "press articles, radio/TV" and other supporting documents optionally produced by the data subjects.

As a premise the Authority observes that the rating processed by the system "could have a heavy impact on the (also private) life of the registered individuals, influencing their choices and prospects and conditioning their very admission to (or exclusion from) specific services, services or benefits";

[26] On data profiling from a constitutional point of view, see the very recent contribution by A. Gatti, *Profilazione e diritti fondamentali*, ES, Napoli, 2025.

[28] As early as 2015, President Antonello Soro argued: "On the basis of these premises, it becomes difficult to implement databases or services aimed at profiling a person - even in moral and relational terms - in order to assess his or her alleged reliability. The European legal order protects privacy, recognized as a fundamental human right, and forbids 'rating' any person in the absence of a public interest, which certainly cannot be invoked when one enters the world of personal and sentimental relationships of anyone who has a social-network account." Antonello Soro, *La reputazione online tra principi di dignità e diritto all'oblio*, Huffington Post, 6 October 2015. On the connection between reputational rating and dignity, see E. di Carpegna Brivio, *Pari dignità sociale e Reputation scoring. Per una lettura costituzionale della società digitale*, Turin, Giappichelli, 2024.

[29] CJEU, Kranemann v Land Nordrhein-Westfalen, 17 March 2005 (ECLI:EU:C:2005:187).

and that the matter must be approached with extreme caution, since "the reputation that [the platform] would like to measure, [...] closely correlated to the consideration of people and to their very projection, social, is intimately connected with their dignity, a keystone of the discipline of personal-data protection".

With reference to the lawfulness of the processing the Authority underlines that the processing was not based on an appropriate source of regulation, since the legal system recognizes and regulates exclusively the "business rating"[30] and not that relating to natural persons. Moreover, the consent of the data subjects, in order to be in accordance with the law, should have been expressed freely, a condition that does not occur in the case of autonomous processing of documents freely usable (e.g. judgments) by an automated system, both with reference to the profiles relating to users who have expressly consented to the processing, and with reference to the "profiles against third parties" relating to subjects not registered on the platform; not to mention that the processing was based on inadequate information.

The processing in question then appeared in violation of the principles of data minimization, since it was based on a massive collection of data and documents, and of proportionality of processing, in that the relevance and pertinence of the data and documents collected appeared doubtful and, in any case, unproven and the period of data retention was not adequately justified. Finally, the processing, involving a potentially very large number of subjects, would have had reliable significant repercussions for the rights of the data subjects in the event of a breach of security measures and the latter were not endowed with sufficient reliability.

The Mevaluate Association challenged the Authority's measure before the Court of Rome which, by judgment no. 5715/2018, held that the lack of regulatory discipline on reputational rating did not prevent the development of infrastructures capable of assigning it. Of the opposite view compared to the Authority, which had considered that the complex system of collection and processing of personal data at issue would have been able to affect both the economic and social representation of a broad category of subjects and the private life of the registered individuals, the ordinary judge noted how private evaluation and certification bodies, recognized also for the purposes of certification of quality and/or conformity to technical standards, are now widely used.

The Court, establishing that "one cannot deny to private autonomy the power to organize accreditation systems of subjects, providing broadly 'evaluative' services, in view of their entry into the market, for the conclusion of contracts and for the management of economic relations," thus considered the processing to be in accordance with the requirement of lawfulness of processing, starting from the assumption that "the activities of uploading information and of validation and certification of the documents are subject to the consent of the data subject and to the voluntariness of his action". On the basis of such arguments, therefore, it annulled the Authority's measure, subject to the prohibition of processing the personal data of subjects not registered on the platform, even if taken from documents that are freely knowable.

With reference, instead, to the profiles against third parties and to the other issues raised by the Authority, the Court confirmed the unlawfulness of the processing carried out by Mevaluate.

On the question of the validity of the consent given by the data subject, the Court of Cassation ultimately ruled, which, in judgment 14381/2021, upheld the Authority's appeal and expressed the following principle of law: "in the matter of processing of personal data, consent is validly given only if it is expressed freely and specifically with reference to clearly identified processing; it follows that in the case of a web platform (with annexed IT archive) preordained to the processing of reputational profiles of individual natural or legal persons, centered on a calculation system with an algorithm at its base aimed at establishing reliability scores, the requirement of awareness cannot be considered satisfied where the executive scheme of the algorithm and the elements of which it is composed remain unknown or not knowable by the data subjects"[31].

In fact, the Supreme Court underlined that for the lawfulness of processing based on consent, Art. 23 of Legislative Decree No. 196 of 2003 (i.e. the pre-GDPR privacy code, in force at the time of the facts as the reference legislation) presupposed not only consent tout court, but also that the consent was validly given[32]. Specifically, indeed, Art. 23 provided that "(a) the processing of personal data by private parties or public economic bodies is permitted only with the express consent of

[30] Managed by the National Anti-Corruption Authority (ANAC) on the basis of Article 83(10) of Legislative Decree No. 50 of 18 April 2016.

[31] Decision of the Italian Data Protection Authority (Garante per la protezione dei dati personali) No. 488 of 24 November 2016 [web doc No. 5796783]

[32] Corte di Cassazione, decisions No. 17278/2018 and No. 16358/2018

the data subject; (b) consent may concern the entire processing or one or more operations of the same; (c) consent is validly given only if it is expressed freely and specifically with reference to 'clearly identified' processing, if it is documented in writing, and if the information referred to in Art. 13 has been given to the data subject; (d) consent is expressed in written form when the processing concerns sensitive data".

In the present case, therefore, the processing was (and is) functional to determining the reputational profile of the subjects and, consequently, the assessment of the lawfulness of such processing, based on consent, could not be envisaged by the court without a prior consideration of the elements likely to affect the seriousness of the manifestation, including the elements implicated and considered in the relevant algorithm, the functioning of which is essential to the calculation of the rating[33].

In particular, the Court of Cassation highlights that the poor transparency of the algorithm with respect to the specific purpose to which it was preordained had not been considered decisive by the appealed judgment. This approach is not shared by the Supreme Court which maintains that the prerequisite of the lawfulness of processing is precisely constituted by the validity of the consent that is assumed to have been given at the time of accession. And it cannot logically be affirmed that the accession to a platform by the members of society also includes the acceptance of an automated system, which makes use of an algorithm, for the objective evaluation of personal data, where the executive scheme in which the algorithm is expressed and the elements considered for the purpose are not made knowable.

As is well known, the observations of the Supreme Court, pertinent and precise, are amply confirmed by the GDPR where Art. 4, paragraph 11, establishes that the consent of the data subject, in an unequivocal manner, must first of all be free because if the data subject does not make a real choice and feels obliged to give his consent also to avoid negative consequences in the event that he refused, then the consent cannot be considered valid. If, for example, consent is inserted in a non-negotiable part of terms and conditions it is presumed that it has not been given freely. Moreover, consent must be specific, since if it is used to justify multiple processing operations it must be freely given for each one. Data subjects must be able to choose for which purposes they consent to processing. Finally, consent must be informed: without accessible information, data subjects cannot make informed decisions and therefore, clarifies the WP29 (in its 2017 guidelines) "user control would become illusory and the consent invalid for the processing".

With reference then to forms of automatic profiling, it must be pointed out that in the field of decisions based solely on automated processing, as provided by the already mentioned Art. 22, the Regulation introduces the need to provide the data subject with more information on the methods of creation and use of these processes. In this regard, indeed, the GDPR finds significant risks for the rights and freedoms of individuals connected to the tendency toward opacity of automated processes and mechanisms, which often leads the individual, the subject of profiling, not to be aware of it and to the creation, by the controller, of new data, additional to the original ones, which could reduce the data subject to a category to which he does not recognize himself, thus conditioning his choices and, in some cases, also leading to forms of discrimination.

Therefore, the GDPR, to correct this informational imbalance between controller and data subject and to avoid prejudice to the legal sphere of the latter, identifies a series of requirements on which to focus in order to make these automated processing operations compliant with the legislation: specific prescriptions on transparency and fairness; greater obligations of accountability; specific legal bases for the legitimization of processing; safeguards for individuals in terms of the right to object to profiling and, in particular, to profiling for marketing purposes; the conduct of a data-protection impact assessment where certain conditions are not met.

In particular, Art. 13, para. 2, letter f) and Art. 15, para. 1, letter h) establish the right of the data subject to know of the existence of automated decision-making and, in particular, to obtain meaningful information about the logic used (the criteria assumed to reach the decision, without thereby necessarily having to provide a complex explanation of the algorithms used) and about the envisaged consequences of such processing (through examples one will have to provide information on how the automated process could in future affect the person concerned).

Taking into account the significant risks to the rights and freedoms of the data subject for these types of processing, the Regulation, on the one hand, obliges the controller to implement appropriate and strengthened protective measures (it will also be important to provide methods that regularly verify the correctness of the processes to limit classification or evaluation errors with a negative impact on the profiled subjects), on the other hand, recognizes the power of the data subject to obtain human intervention on the part of the controller, to express his opinion and to contest the

[33] F. Galli, *Rating reputazionale e diritto alla spiegazione,* in *Labour & Law Issues*, vol. 9, No. 2, 2023, pp. 62-97.

decision, in cases where such decision is provided for by contract or consented to by the data subject (Art. 22, para. 3). Among other things, an automated decision-making process involving special categories of data, as per Art. 9, para. 1, is permitted only in the presence of the explicit consent of the data subject or for reasons of important public interest on the basis of Union or Member State law.

It is worth recalling, in fact, what is established by Art. 22, para. 1 GDPR: "the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her".

For the Group of European Data Protection Authorities[34], therefore, Art. 22, para. 1, establishes a prohibition for that individual decision-making process that is completely automated, including profiling, which has a legal or similar effect on the data subject. By "decision based solely on automated processing" one must understand a decision taken without the involvement of a human being who can influence and possibly change the result through his authority or competence.

For the right of the data subject not to be subject to a decision based solely on automated processing to be recognized, it is necessary that such decision "produces legal effects or similarly significantly affects his person". The reference to "legal effects" concerns, obviously, the impact that an automated decision can produce on the legal sphere of the individual (e.g. penalizing the right of association, of voting, of contractual freedom, of free movement, etc.). In addition to this, the rule also opens up to circumstances that "in a similar way" can potentially and significantly influence the behaviors and choices of the individuals concerned.

Recital 71 of the GDPR cites, as examples of automated decisions that can significantly affect the rights and freedoms of individuals, the automatic refusal of an online credit application or electronic hiring practices without human interventions.

Art. 22 in para. 2 provides three exceptions to the general prohibition of a completely automated decision-making process that brings effects in the legal sphere of the individual: when the decision is necessary for the conclusion or performance of a contract between the data subject and a controller; when the decision is authorized by the law of the Union or of the Member State to which the controller is subject; when the decision is based on the explicit consent of the data subject.

With reference to the first point, the European Authorities clarify that the necessity to use automated decisions for the performance or conclusion of a contract must be interpreted restrictively: this means that the controller must be able to demonstrate that the profiling is necessary (principle of necessity) and that no less invasive alternative means are available (principle of proportionality), along the lines of what has already been established by the ECtHR case law, with particular reference to Art. 8 of the Convention[35]. With reference to the second point, the legislation of the Member States can, in specific cases, authorize the use of an automated decision-making process for the monitoring and prevention of fraud and tax evasion or to guarantee the security and reliability of a service provided by the controller. Finally, as regards the third derogation, the Regulation requires the explicit consent of the data subject, that is, confirmed by an express declaration and not inferred from conclusive conduct.

On the administrative front, one must finally not forget that the Council of State[36], too, in assessing the legitimacy of the adoption of an algorithm for carrying out an administrative activity, has noted that three fundamental principles must be duly taken into consideration in the examination and use of IT tools. First, the principle of knowability, by which everyone has the right to know of the existence of automated decision-making processes that concern him and, in this case, to receive meaningful information about the logic used. The second principle can be defined as the principle of non-exclusivity of the algorithmic decision. In the case in which an automated decision "produces legal effects that concern or significantly affect a person," that person has the right that such decision is not based solely on such automated process. The third principle is that of algorithmic non-discrimination, according to which it is appropriate that the controller use appropriate mathematical or statistical procedures for profiling, implementing adequate technical and organizational measures in order to ensure, in particular, that the factors that entail inaccuracies in the data are corrected and that the risk of errors is minimized.

[34] EDPB, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, 22 August 2018.

[35] S. & Marper v United Kingdom, ECtHR, apps. nos. 30562 and 30566, 4 December 2008, §§ 35, 119.

[36] Consiglio di Stato, decision No. 8472/2019.

### 4.2. The school context

Also to be recalled on the subject is the question of reputational rating in the school context.

The Data Protection Authority sent a request for information to the association Crop News (where "Crop" is an acronym for Personalized Objective Reputational Chronicles, which operates within the Mevaluate group). Press reports revealed that the association, within the framework of the "Virtute 4 Students" Project, was experimenting on students the reputational rating processed on the basis of algorithms by the Mevaluate Platform itself and that a higher education institution had joined the project.

The Garante, considering the delicacy of the project which is addressed to particularly vulnerable subjects (students and minors), asked the association for clarifications, in particular the functioning of the platform and of the connected database in order to allow the Authority to evaluate the impact of the use of algorithms and the effects that they can determine on students, as well as the measures possibly adopted for their protection[37].

The Association declared that the purpose of the initiative was to stimulate students to build a real reputation that coincides with the virtual reputation through their own digitized reputational rating in order to promote a transparent behavioral economic model. Virtute calculates the individual reputations published by the online periodical Crop News, thus promoting lawful behaviors founded on the fear of appearing in the newspaper. The reduction of unlawful acts and contractual defaults would guarantee both economic operators and consumers and users, but by reason of the moral "blackmail" of an automated reputational rating, which moreover assigns variable remuneration in relation to the contribution provided by each person to qualify the documented and traceable reputation of oneself and of the counterparties in obligatory relationships.

On the basis of the information provided and of further investigations, the Authority considered the processing unlawful, in violation of the mandatory safeguards of Art. 22 of the GDPR and, above all, contrary to the principles of lawfulness (since it was carried out in the absence of a sufficient legal basis), fairness (since it is capable of engendering unreasonable and disproportionate discriminations and prejudices), transparency (in that it was carried out in the absence of adequate information notice), minimization (processing quantities of data that are indeterminate and not determinable a priori), limitation (since it produces new data that can be processed for the pursuit of purposes that are different and not necessarily lawful), and storage limitation (entrusting the data to an online newspaper). Crop News was therefore admonished, while a new investigation was opened against Mevaluate, collector of all the data collected.

## 5. Concluding reflections. From algocracy to algoretics

The experiences and examples shown in the course of the contribution allow us an overview of the practical realizations, mostly experimental, that to date concern reputational profiling. However, there is no need to say that, because of its implications, the subject needs to be addressed at a higher level, that is, at the level of its theoretical and philosophical foundations that lead us to come to terms with the phenomenon of the so-called algocracy, that is, the increasing dependence on algorithms and automated decision-making processes in various aspects of society. This tendency has been driven by the progress of artificial intelligence and machine learning, which have made it possible to automate activities that previously were considered the exclusive domain of human judgment and skills. However, since the use of algorithms has become more widespread, many concerns have arisen about the ethical implications of these systems and their impact on individuals and on society as a whole.

In response to these questions, we have witnessed a shift of attention to algoretics, that is, the study of the ethical implications of algorithms and automated decision-making processes. Algoretics seeks to address issues such as the potential for algorithms to perpetuate and amplify existing inequalities and prejudices, as well as the lack of transparency and accountability in these systems. Overall, the shift from algocracy to algoretics reflects a growing recognition of the need to consider the ethical implications of algorithms and automated decision-making systems. Although these technologies have the potential to improve efficiency and objectivity in decision-making, it is important to ensure that they are used in a way that promotes fairness and transparency. This will require constant attention to the ethical and social implications of algocracy and a commitment to addressing any negative consequences that might arise.

[37] "*Scuola: rating reputazionale sotto la lente del Garante privacy. Avviata istruttoria su una piattaforma rivolta agli studenti*", Garante press release, 3 May 2022.

Machine learning, AI and Big Data are undoubtedly changing the way of reading and managing reality. These technologies not only help man by increasing his capabilities, but in an ever greater number of situations give rise to systems, bots or robots, completely autonomous, which therefore replace him.

In particular, artificial intelligences are technological artifacts different from all the artifacts produced up to today. All the tools that we have contrived have enabled man to perform certain tasks by enhancing his physical capacities: from primitive clubs up to large industrial machines, all these tools have served to perform specific tasks better, faster and more effectively. AIs, both in bots and in robots, go beyond the concept known up to now. All the automatic mechanisms built during the industrial revolution were designed according to what their purpose was and, therefore, performed exclusively that specific task.

Today AIs are not designed like this: they are not programmed software, but trained systems, surpassing the classic "if this then that" model in which an engineer first foresaw all the possible occurrences. AIs respond autonomously to a problem that is posed to them. These artifacts are a new species among machines. Of machine sapiens that share the world with homo sapiens, however reasoning differently.

Data scientists underline how the problem is linked to the quality and to the quantity of the data that the machines have at their disposal to make their decisions: when we have a perfect database on which to run AI systems - so they say - the machine will make "perfect" choices.

But is it really so? Already in the past we had this impression. Laplace maintained that if we had known the position at the same instant of all the particles that contain the universe we would have been able to predict all the future and to know all the past of the universe. But even if we were able to create a map that is the exact copy of reality, including within it everything, this map would still be useless, since it would prove to be as complex as reality and, therefore, too complex to make decisions more effective than those that we already make.

We would thus find ourselves before the well-known paradox recounted by Jorge Luis Borges in a fragment On Exactitude in Science, the last of A Universal History of Infamy first published in 1935. As is his habit, the Argentine author attributes the quotation to a book that in reality does not exist: "... In that Empire, the art of cartography attained such perfection that the map of a single Province occupied an entire city, and the map of the empire an entire Province. Over time, these unwieldy maps were no longer sufficient. The colleges of cartographers made a map of the Empire that had the vastness of the Empire and coincided perfectly with it (Suárez Miranda, *Viajes de varones prudentes*, book IV, chap. XIV, Lérida, 1658)".

Data are a map of reality, they represent a reduction of reality and for this they are useful for making decisions. Moreover AIs, besides databases, work on sensors which in turn are not able to read all reality: they learn only a part of it transforming it into data. This is the key point of the question. Since artificial intelligences base their decisions on data, and since these are not a perfect copy of reality, it cannot be thought a priori that the machine endowed with artificial intelligence can make a choice devoid of errors. The machine sapiens will always and constitutively be fallible.

Since artificial intelligences can make mistakes, it is necessary to understand how to manage such errors, assigning to AIs an ethics. That is, it is necessary to find a shared ethical system so that the use of these systems does not produce injustices, does not harm people and does not create strong global imbalances.

The existence of machine sapiens calls for setting up a new universal language that knows how to translate these ethical guidelines into directives executable by the machine and since the dimension of the digital era is regulated by algorithms, here algocracy imposes itself, the dominion of the algorithm, in the face of which it is urgent to develop the common language of algoretics.

To the extent that we want to entrust human skills of understanding, judgment and autonomy of action to AI software systems we must understand the value, in terms of knowledge and capacity for action, of these systems that claim to be intelligent and cognitive, and it is for this reason that the problem is first and foremost philosophical and epistemological. AIs "work" according to schemes that connect data; but what kind of knowledge is this? What value does it have? How should it be treated and considered? It is first of all necessary to clarify in what sense one speaks of value. In fact algorithms work on values of a numerical nature, while ethics weighs moral value. It would therefore be necessary to establish a language that knows how to translate moral value into something computable by the machine. The perception of ethical value is a purely human capacity and the capacity to work with numerical values is instead the ability of the machine. But in the relationship between man and machine the true knower and bearer of value is the human part. The treatment on human dignity and human rights can only teach that it is man who must be protected in the relationship between man and machine.

This evidence provides us with the fundamental ethical imperative for the machine sapiens: doubt itself. This means having to put the machine in a position to nurture a certain sense of uncertainty. Every time the machine does not know whether it is safeguarding the human value with certainty it must request the action of man. This fundamental directive is obtained by introducing statistical paradigms within AIs. This capacity for uncertainty must be the heart of the machine's deciding, since by doing so the machine itself places the human at the center.

## References


Ahl, B., L. Catá Backer, y Y. Chen. "Law and Social Credit in China." *China Review* 24, no. 3 (2024): 1–15

Angwin, J. et al. "Machine Bias." *ProPublica*, May 23, 2016.

Arendt, Hannah. *The Origins of Totalitarianism*. New York: Schocken Books, 1951.

Autoriteit Persoonsgegevens. "Belastingdienst beloofde ambtenaren niet te straffen om toeslagenaffaire." May 20, 2019.

Autoriteit Persoonsgegevens. "Werkwijze Belastingdienst in strijd met de wet en discriminerend." July 17, 2020.

Castro, C. "What's Wrong with Machine Bias." *Ergo* 5, no. 15 (2019–2020).

Chin, J., and L. Lin. *Surveillance State*. New York: St. Martin's Press, 2022.

"China's Chilling Social Credit Blacklist." *Human Rights Watch*, December 12, 2017.

"China Releases Investigative Journalist After Almost Year in Jail." *Newsweek*, August 3, 2014.

Creemers, R. *China's Social Credit System: An Evolving Practice of Control*. SSRN, 2018.

Dai, X. "Toward a Reputation State: A Comprehensive View of China's Social Credit System Project." In *Social Credit Rating*, edited by O. Everling, 139–165. Wiesbaden: Springer, 2020.

Decision of the Italian Data Protection Authority (Garante per la protezione dei dati personali) No. 488 of November 24, 2016 [web doc No. 5796783].

Di Carpegna Brivio, E. *Pari dignità sociale e reputation scoring: per una lettura costituzionale della società digitale*. Turin: Giappichelli, 2024.

European Data Protection Board (EDPB). *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*. August 22, 2018.

Elliott, A. Fenwick. "China Is Banning People with Bad 'Social Credit' from Using Planes and Trains." *The Telegraph*, March 19, 2018.

Fenger, M., and R. Simonse. "The Implosion of the Dutch Surveillance Welfare State." *Social Policy & Administration* 58, no. 2 (2024): 264–276.

Galli, F. "Rating reputazionale e diritto alla spiegazione." *Labour & Law Issues* 9, no. 2 (2023): 62–97.

Garante per la protezione dei dati personali. "Scuola: rating reputazionale sotto la lente del Garante privacy. Avviata istruttoria su una piattaforma rivolta agli studenti." Press release, May 3, 2022.

Gatti, A. "L'algoritmo tra volontà e rappresentazione." *DPCE Online*, no. 3 (2020): 3457–3461.

Gatti, A. *Profilazione e diritti fondamentali*. Napoli: ES, 2025.

Hofmann, S. *Social Credit: Technology-Enhanced Authoritarian Control with Global Consequences*. International Cyber Policy Center, Policy Brief Report no. 6, 2018.

Johnson, S. "How China's 'Social Credit Score' Will Punish and Reward Citizens." *Big Think*, April 26, 2018.

Lin, L. Yu-Hsin, and C. Milhaupt. "China's Corporate Social Credit System: The Dawn of Surveillance State Capitalism?" *The China Quarterly* 256 (2023): 835–853.

Liu, C., and A. Rona Tas. "Trusting by Numbers: An Analysis of a Chinese Social Credit System Governance Infrastructure." *Critical Sociology* 51, no. 6 (September 2025): 1247–1265.

Loefflad, C., M. Chen, and H. Grossklags. "Reputational Discrimination and Fairness in China's Social Credit System." *Digital Government* 5, no. 4 (December 2024): 1–27.

Mac Sithing, D., and M. Siems. "The Chinese Social Credit System: A Model for Other Countries." *Modern Law Review* 82, no. 6 (2019): 1034–1071.

National Development and Reform Commission. *Social Credit Action Plan 2024–2025*. June 2024. https://www.chinalawtranslate.com/en/2024-2025social-credit-plan/

Parlementaire ondervragingscommissie Kinderopvangtoeslag. *Tweede Kamer der Staten-Generaal*. July 11, 2020.

Soro, Antonello. "La reputazione online tra principi di dignità e diritto all'oblio." *Huffington Post*, October 6, 2015.

Stradella, E. "Stereotipi e discriminazioni: dall'intelligenza umana all'intelligenza artificiale." *Consulta Online,* March 20, 2020. www.giurcost.org.

"Uitspraak Bulgarenfraude: hoe zat het ook alweer?" *RTL Nieuws*, May 19, 2015.

van Blomberg, M. "The Social Credit System in China's Rule of Law." *Mapping China Journal*, no. 2 (2018): 78–162.

Venice Commission, Council of Europe. *Opinion No. 1030/2021*.

Zuo, Z. *Governance by Algorithm: China's Social Credit System*. University of Cambridge Working Paper, June 16, 2020.