

https://dspace.org/

Artificial intelligence and automated decision making under the GDPR and the POPIA

Goldman, Gavin David

2024
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/43523
Goldman, Gavin David . (2024). Artificial intelligence and automated decision making under the GDPR and the POPIA [Master's dissertation, University of the Witwatersrand, Johannesburg].WireDSpace.

Downloaded from a DSpace repository. https://dspace.org/

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG | WITS SCHOOL OF LAW

ARTIFICIAL INTELLIGENCE AND AUTOMATED DECISION MAKING UNDER THE GDPR AND THE POPIA

Gavin David Goldman

Supervisor: Prof. E Zitzke

Submitted in partial fulfilment of the requirements for the degree of

Master of Laws by Coursework and Research Report

at the University of the Witwatersrand, Johannesburg

Date: 9 May 2024

DECLARATION

I declare that this report is my own, unaided work. It is submitted in partial fulfilment of the requirements for the degree of Master of Laws in the field of Information and Communications Law at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any other degree or examination in any other university.

_____

8320036

30 SEPTEMBER 2023

ABSTRACT

This analysis considers the concepts of AI and machine learning and examines their reliance on the processing of personal data and the challenges this poses from a data-privacy and human-rights perspective, particularly in relation to profiling. It evaluates the effectiveness of the General Data Protection Regulation (GDPR) and the Promotion of Personal Information Act 4 of 2013 (POPIA) in regulating Automated Decision Making (ADM) and considers the limitations of the right to an explanation under these provisions. The analysis proposes that the current framework of the GDPR and POPIA does not clearly address the issue of explainability and that the focus should shift to providing a data subject with a counterfactual to give practical effect to this right which would better serve data subjects.

Table of Contents

DECLARATION 1

ABSTRACT 2

I. INTRODUCTION 4

II. AI AND MACHINE LEARNING 5

(a) Supervised Learning 6

(b) Unsupervised Learning 7

(c) Reinforcement Learning 7

III. AUTOMATED DECISION MAKING (ADM) 7

IV. PROFILING 9

V. THE GDPR AND THE POPIA 10

VI. THE RIGHT TO AN EXPLANATION 23

VII. THE CHALLENGE 26

VIII. THE WAY FORWARD 27

IX. CONCLUSION 30

BIBLIOGRAPHY 31

I. INTRODUCTION

Human history has been characterized by innovation and technological advances. Most notably, the meteorological rise in computing power has had a knock-on effect on many technological milestones. Traditionally, the use of computers has aided users in undertaking tasks that could never have been accomplished before. With the advent of Artificial Intelligence (AI), complicated and routine tasks and processes can be accomplished with very little or no human intervention. Kroll, Barocas, Felton, Reidenberg, Robinson & Yu state the position succinctly:

> 'Many important decisions that were historically made by people are now made by computer systems: votes are counted; voter rolls are purged; loan and credit card applications are approved; welfare and financial aid decisions are made; taxpayers are chosen for audits; citizens or neighborhoods are targeted for police scrutiny; air travelers are selected for search; and visas are granted or denied.'[1]

The influence of AI is prolific in other spheres as well. It has, for example, been instrumental in determining the content of news feeds, promoting targeted advertising, personalizing search results and enabling the use of digital personal assistants such as Siri and Alexa.[2] Importantly, AI has also been instrumental in the development of automated decision making (ADM) which, in simple terms, refers to 'decisions made by technological means'.[3]

The success of AI is inextricably linked to the availability of large volumes of data and the capability of hardware systems to process and interpret these significant volumes of information.[4]

The challenge presented by AI is to balance the ability of developers to innovate using AI with an individual's right to privacy and to be treated fairly without discrimination or bias. Shackelford & Dockery state:

---

[1] Joshua A Kroll, Solon Barocas, Edward W Felton, Joel R Reidenberg, David G Robinson & Harlan Yu 'Accountable Algorithms' (2017) 165 *University of Pennsylvania Law Review* 633 at 636.
[2] Clarke D Asay 'Artificial Stupidity' (2020) 61 *William & Mary Law Review* 1187 at 1191.
[3] Ella Brownlie 'Encoding Inequality: The Case for Greater Regulation of Artificial Intelligence and Automated Decision-Making in New Zealand' (2020) 51 *Victoria University of Wellington Law Review* 1 at 2.
[4] Stuart J Russell & Peter Norvig *Artificial Intelligence A Modern Approach* 3 ed (2009) 27-28.

> 'Because of the substantial conveniences, economic opportunities, and national security challenges related to AI, the need for AI governance is manifest. Some experts caution that regulation will stifle innovation and should be approached carefully or not at all while others argue that the potential malicious uses of general AI should stop its development entirely.'[5]

In part II of this analysis, I will consider what encompasses AI and machine learning and the way in which these technologies rely on the processing of personal data. In Part III, I will also consider how AI and machine learning have enabled the increased use of ADM and the specific challenges this has created, particularly in the context of profiling which is specifically addressed in Part IV. In Part V, I consider how the GDPR and by extension, the POPIA, regulates ADM and whether these mechanisms are likely to be effective. From a South African perspective, the analysis, debate and case law developed under Article 22 of the GDPR, specifically in relation to ADM and the nature and scope of a right to an explanation, is particularly important in relation to the future application and development of s 71 of the POPIA since these issues have not yet been considered here in any meaningful way by the academic community and our courts. The controversial right to an explanation is discussed in Part VI and the challenges associated with it, in Part VII. Finally, in Part VIII, I offer some suggestions on the way forward.

II. AI AND MACHINE LEARNING

There is no standard definition of AI. Katyal formulates the following definition, which for the purpose of this analysis, is appropriate. She describes AI as comprising:

> 'machines that receive inputs from the environment, then learn from or interpret those inputs, and then potentially take certain actions or decisions that affect the environment.'[6]

AI is broadly categorized as being either general or narrow. General AI contemplates a system of intelligence similar to that of a human which can adapt to changes in circumstances 'to generalize knowledge from one goal or context to others'.[7] This type of

---

[5] Scott J Shackelford & Rachel Dockery 'Governing AI' (2020) 30 *Cornell Journal of Law & Public Policy* 279 at 300.
[6] Sonia K Katyal 'Private Accountability in the Age of Artificial Intelligence' (2019) 66 *UCLA Law Review* 54 at 62.
[7] Ben Goertzel 'Artificial General Intelligence: Concept, State of the Art, and Future Prospects' (2014) *Journal of Artificial General Intelligence* 5(1) 1 at 1.

AI, however, has not, as yet, been achieved and cannot match 'higher-order human abilities' at this stage.[8]

Narrow AI, is directed towards a specific task and cannot exhibit any degree of intelligence beyond that task without reprogramming.[9] Examples of narrow AI include voice assistants, smart devices, wi-fi doorbells and cybersecurity systems.[10]

An AI system relies on data in order to function. Programming that prescribes a series of instructions that must be followed for the relevant software to function is often referred to as an algorithm which, stripped to its bare essentials, is a set of rules which are prescribed in order to achieve a particular task, much like a cooking or baking recipe.[11] What distinguishes AI from any other system is that it does not need any human intervention to function.[12] This ability is achieved through machine learning involving algorithms that are able to learn or enhance performance in relation to a particular activity.[13] The most distinct characteristic of machine learning is the computer's ability to improve its performance without the involvement of a human.[14]

There are three main approaches to machine learning that are relevant in the context of AI: supervised learning, unsupervised learning, and reinforcement learning.

(a) Supervised Learning

A supervised machine learning algorithm uses 'labelled' training data sets to classify '"objects", a catchall concept that can include anything, including people'.[15] A simple example would entail the classification of animals between dogs and cats.[16]

---

[8] Harry Surden 'Artificial Intelligence and Law: An Overview' (2019) 35 *Georgia State University Law Review* 1305 at 1309.
[9] Goertzel op cit note 7 at 1.
[10] Ashish Ghosh, Debasrita Chakraborty & Anwesha Law 'Artificial intelligence in Internet of things' (2018) 3 *CAAI Transactions on Intelligence Technology* 208 at 211-214.
[11] Maya C Jackson 'Artificial Intelligence & Algorithmic Bias: The Issues With Technology Reflecting History & Humans' (2021) 16 *J.Bus & Tech* 299 at 301.
[12] Victor M Palace 'What if Artificial Intelligence Wrote This: Artificial Intelligence and Copyright Law' (2019) 71 *Florida Law Review* 217 at 218.
[13] Harry Surden 'Machine Learning and Law' (2014) 89 *Washington Law Review* 87 at 88.
[14] Erik Brynjolfsson & Andrew McAfee 'The Business of Artificial Intelligence' (2021) *Harvard Business Review Special Issue* 20 at 22.
[15] Michael L Rich 'Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment' (2016) 164 *University of Pennsylvania Law Review* 871 at 881.
[16] Azziz Z Huq 'Constitutional Rights in the Machine-Learning State' (2020) 105 *Cornell Law Review* 1875 at 1886.

(b) Unsupervised Learning

Unsupervised learning systems work with unlabelled data and 'infers information from the data set and can be highly resource intensive, as the data set is tested against a massive number of potential patterns'.[17]

(c) Reinforcement Learning

A reinforcement learning system learns through 'trial-and-error'[18] and is able to modify its responses depending on the reaction it obtains.[19] These manifestations of machine learning are important to the extent that they feature prominently in the context of ADM and can arrive at outcomes that may perpetuate certain biases or discriminatory conduct.[20]

III. AUTOMATED DECISION MAKING (ADM)

A decision is said to be automated when it has been arrived at in the absence of any human intervention.[21] It has also been suggested the term addresses the use of algorithms to assist in human decision-making as well as the automated implementation of decisions.[22]

A shortcoming of machine learning, and in particular, supervised learning, is highlighted in the study by Ribeiro, Singh and Guestrin[23] in which the researchers sought to demonstrate how unintended correlations can take place during the course of learning. In the experiment, a machine that was trained to differentiate Wolves (against snowy backgrounds) from Huskies (without snow) misclassified a Wolf without snow as a Husky and a Husky with snow as a Wolf. This study is of particular significance in the context of ADM in circumstances 'when variables used in machine learning models to make predictions about

---

[17] Kristen Johnson, Frank Pasquale & Jennifer Chapman 'Artificial Intelligence, Machine Learning and Bias in Finance: Toward Responsible Innovation' (2019) 88 *Fordham Law Review* 499 at 506.
[18] Jess Whittlestone, Kai Arulkumaran & Matthew Crosby 'The Societal Implications of Deep Reinforcement Learning' (2021) 70 *Journal of Artificial Intelligence Research* 1003 at 1006.
[19] Ibid.
[20] Robin Allen & Dee Masters 'Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making' (2020) 20 *ERA Forum* 585 at 588.
[21] Ibid at 586.
[22] Elvin Evrim Dalkilic 'A Way Out For Europe: How can Europe Combat Discrimination by Automated Decision Systems' (2022) 9 *Lincoln Memorial University Law Review* 37 at 38.
[23] Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin 'Why Should I Trust You?': Explaining the Predictions of Any Classifier' *arXiv* (2016), available at: *http://arxiv.org/abs/1602.04938*, accessed on 24 August 2022.

or classify people - such as qualifications, work history, or criminal record – are already shaped by structural discrimination'.[24]

This shortcoming was identified in a study by ProPublica on the COMPAS system, an algorithm that was meant to predict the probability of defendants' recidivism.[25] The system demonstrated bias favouring white over black defendants due to its dataset largely comprising black defendants.[26] This study, however, has not been universally accepted and a number of contrary views as to its findings have been expressed.[27]

Despite criticisms of the ProPublica study, it has been recognised that algorithmic bias has particular adverse implications for the underprivileged and marginalized members of society.[28] Gravett states that the current debate about algorithmic bias suggests that it may not only be a 'possibility' but 'perhaps even' an 'inevitability'.[29]

The occurrence of bias, however, may not necessarily stem from human bias to the extent that the effectiveness of an algorithmic system will also be influenced by the quality of the data it relies upon.[30] In this context, since machine learning systems are trained on data, the outcome of the relevant prediction may be biased, if the training data is not representative of the broader population being investigated.[31] Human bias, on the other hand, will manifest itself where the individuals involved in the collection of data will, either intentionally or unintentionally reflect their own biases in the compilation of the relevant training data.[32]

An inherent characteristic of a decision-making process driven by an algorithm, is the systems' lack of transparency (opacity) that underlies the relevant decision. Rich makes the point that models based on machine learning can be so complicated that the designers of the algorithm do not know 'how or why the generated model creates accurate predictions'.[33] Lu

---

[24] Reuben Binns 'Human Judgment in algorithmic loops: Individual justice and automated decision-making' (2022) 16 *Regulation & Governance* 197 at 204.
[25] Ibid.
[26] Ibid.
[27] Anne L Washington 'How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica debate' (2019) 17 *The Colorado Technology Law Journal* 1 at 23.
[28] Peter K Yu 'The Algorithmic Divide and Equality in the Age of Artificial Intelligence' (2020) 72 *Florida Law Review* 331 at 355.
[29] Willem Gravett 'Bias and lack of accuracy in risk-assessment software in the United States criminal justice system' (2021) 34 *South African Journal of Criminal Justice* 31 at 33.
[30] Ari Ezra Waldman 'Power and Automated Decision-Making' (2019) 88 *Fordham Law Review* 613 at 622.
[31] Katyal op cit note 6 at 70.
[32] Jackson op cit note 11 at 309.
[33] Rich op cit note 15 at 880.

states the position more broadly that '[a]lgorithmic opacity refers to the technical complexity, trade secrecy, and managerial invisibility of AI systems that hamper stakeholder inspection of the inner workings and operating results of AI'.[34] These opaque systems are referred to as 'black box' models.[35] The challenge stems from the fact that the specifics of neural networks and deep-learning applications are so deeply hidden within the AI system that the achievement of transparency is almost unattainable.[36]

The opacity of black box models is a particularly relevant consideration in relation to an individual's right to transparency and the right to an explanation, which arises in the context of ADM, an issue which is addressed later. The opacity of ADM also has significant implications when individuals who have been subjected to profiling, are unable to understand the decision-making process.

IV. PROFILING

Profiling 'involves collecting information about a person and assessing their characteristics or patterns of behaviour in order to place them in a particular category or group and to draw on that inference or prediction.'[37]

A more precise definition of profiling in the context of AI and decision-making has been offered:

> 'Profiling is a technique of (partly) automated processing of personal and/or non-personal data, aimed at producing knowledge by inferring correlations from data in the form of profiles that can subsequently be applied as a basis for decision-making'.[38]

---

[34] Sylvia Lu 'Data Privacy, Human Rights and Algorithmic Opacity' (2022) 110 *California Law Review* 2087 at 2098.
[35] Cynthia Rudin & Joanna Radin 'Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition' (2019) 1 *Harvard Data Science Review* 3 at 3.
[36] Charlotte A Tschider 'Beyond the "Black Box"' (2021) 98 *Denver Law Review* 683 at 689.
[37] Alessandra Silveira 'Automated individual decision-making and profiling [on Case C-634/21 – SCHUFA (SCHUFA (Scoring)]' (2023) 8 *EU Law Journal* 74 at 75.
[38] Francesca Bosco, Niklas Creemers & Valeria Farraris et al 'Profiling Technologies and Fundamental Rights and Values: Regulatory Challenges and Perspectives from European Data Protection Authorities' in Serge Gutwith, Ronald Leenes & Paul de Hert et al (eds) (2015) *Reforming European Data Protection Law. Law, Governance and Technology Series* 3–33.

There is some overlap between profiling and ADM, however, some hold the view that ADM can take place with or in the absence of profiling.[39] Wiedemann, for example, adopts a strict 'two-step' approach by treating profiling as distinct and separate from the decision-making process which he maintains provides more certainty and the ability to differentiate more clearly between the different legal consequences related to the two processes.[40]

The dangers of profiling are readily apparent. Three risks have been identified[41]: Firstly, the generation of 'new or unknown' information through profiling is 'privacy invasive'. Secondly, the notion of informed consent is challenged when profiling takes place. Thirdly, profiling may result in the incorrect identification, characterization and evaluation of individuals. When these inaccuracies filter through to the making of a decision, there is the potential for harm.[42] It is precisely these risks that the GDPR and the POPIA seek to address.

V. THE GDPR AND THE POPIA

The GDPR came into effect on 25 May 2018. One of its primary objectives, as set out in Article 1, is the 'protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data'.

Article 4 of the GDPR defines 'profiling' as:

> '[a]ny form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements'.

Article 22(1) of the GDPR provides:

---

[39] Article 29 Data Protection Working Party 'Guidelines on Automated Individual Decision Making and Profiling for the Purpose of Regulation 2016/679', 6 February 2018, available at: *https://ec.europa.eu/newsroom/article29/items/612053* accessed on 27 August 2022.
[40] Klaus Wiedemann 'Profiling and (automated) decision-making under the GDPR: A two-step approach' (2022) 45 *Computer Law & Security Review* 1 at 2.
[41] Frederike Kaltheuner & Elettra Bietti 'Data is power: Towards additional guidance on profiling and automated decision-making in the GDPR' (2018) 2 *Journal of Information Rights, Policy and Practice* 1 at 7.
[42] Ibid.

> '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.'

The protection afforded by Article 22(1) will not assist a data subject who has simply been profiled without having been subject to a decision, whether automated or not.[43] In these circumstances, the 'safeguards'[44] provided for in Articles 13 to 15 of the GDPR would be available which cater for the furnishing of information by the controller to a data subject[45] and the right of a data subject to request information from a controller in relation, inter alia, to the purpose of the relevant processing.

Articles 13(2)(f), 14(2)(g) and 15(1)(h) provide for the furnishing of information specifically relating to:

> 'the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.'

It is immediately apparent, however, that the ability to be furnished with or to request information in relation to the existence of ADM is not dependent upon whether there was a decision that was specifically directed towards profiling. It has been recognised that there is some ambiguity between Article 22 and Articles 13 to 15 in the sense that:

> 'the wording in Article 22(1) suggests that profiling is distinct from decision-making and is a form of automated processing while the wording in Articles 13(2)(f), 14(2)(g) and 15(1)(h) "automated decision-making, including profiling" suggests that profiling is itself a form of decision-making'.[46]

In either case, however, the relevant decision need not be associated with profiling for the purposes of Articles 13 to 15 of the GDPR. In other words, a data subject will be entitled to

---

[43] Kaltheuner & Bietti, op cit note 41 at 10.
[44] Ibid.
[45] Articles 13 and 14 are almost identical, the only difference being whether the personal data was collected from the data subject or not.
[46] Kaltheuner & Bietti op cit note 41 at 10.

information in relation to ADM regardless of whether such ADM entailed or resulted in profiling.

Article 22(2) provides for three exceptions to the right contained in Article 22(1) which will not apply if the relevant decision:

> '(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
>
> (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
>
> (c) is based on the data subject's explicit consent'.

The Article 29 Data Protection Working Party which was tasked to provide guidelines[47] on Article 22 points out that Article 22 will not only apply when it is 'actively invoked' by a data subject but that it operates as a 'general prohibition' against decision making 'based solely on automated processing'.[48] The Working Party, however, reiterates that the prohibition specifically applies only when a decision, based soley on automated processing including profiling, has a legal effect on someone or affects them in a similarly significant manner.[49] There are, however, the prescribed exceptions which permit processing of the nature contemplated in Article 22(1) even where the aforementioned effects are present.[50]

The POPIA's objectives are similar to those of the GDPR. It came into effect on 1 July 2021 and its purpose, in very brief terms, as set out in section 2, is the 'safeguarding of personal information' to give 'effect to the constitutional right to privacy' and to 'regulate the manner in which personal information may be processed'.

Section 71(1) of the POPIA provides:

---

[47] Brownlie op cit note 3. There are no equivalent guidelines that have been published in relation to the application of the POPIA. The Article 29 Working Party's guidelines are nonetheless instructive in highlighting some of the issues associated with ADM which will have some relevance in the context of the POPIA.
[48] Article 29 Data Protection Working Party op cit note 39 at 19.
[49] Ibid at 20.
[50] Ibid.

> ‘Subject to subsection (2), a data subject may not be subject to a decision which results in legal consequences for him, her or it, or which affects him, her or it to a substantial degree, which is based solely on the basis of the automated processing of personal information intended to provide a profile of such person including his or her performance at work, or his, her or its credit worthiness, reliability, location, health, personal preferences or conduct.’

Like Article 22 of the GDPR, s 71 is similarly restricted to circumstances which culminate in a decision so that profiling alone, a term which is not specifically defined, will not fall under its provisions.

Similar safeguards, however, to those provided for in Articles 13 to 15 of the GDPR are to be found in sections 18 and 23 of the POPIA, except that those sections do not cater specifically for the furnishing of information in relation to the existence of ADM but rather cater for the furnishing of information by a responsible party to a data subject in relation to the collection of personal information and the data subject’s right of access to such information.

In a somewhat cumbersome fashion, instead of housing the prohibition or the right against automated decision-making within s 71 itself, s 5(g) of the POPIA provides that:

> ‘a data subject has the right to have his, her or its personal information processed in accordance with the conditions for the lawful processing of personal information as referred to in Chapter 3, including the right –
>
> (g) not to be subject, under certain circumstances, to a decision which is based solely on the basis of the automated processing of his, her or its personal information intended to provide a profile of such person as provided for in terms of section 71’.

Section 71(2) provides for the following exceptions to sub-section (1):

> ‘The provisions of subsection (1) do not apply if the decision—
>
> (a) has been taken in connection with the conclusion or execution of a contract, and—
>
> (i) the request of the data subject in terms of the contract has been met; or
>
> (ii) appropriate measures have been taken to protect the data subject’s legitimate interests; or

(b) is governed by a law or code of conduct in which appropriate measures are specified for protecting the legitimate interests of data subjects.'

A somewhat peculiar omission from the list of exceptions provided for in s 71(2) is where the decision is based on the explicit consent of a data subject. What makes this omission strange is that s 69 of the POPIA, for example, which deals with the prohibition against direct marketing by means of unsolicited electronic communications, includes the giving of consent by the data subject as one of the exceptions to the prohibition. It is therefore all the more confusing why consent should also not feature as one of the exceptions included in s 71(2).

Article 22 and s 71 are unduly complicated, convoluted, and confusing. It is difficult to predict how a controller or responsible party, not to mention the average data subject, will be able to navigate through these series of loops, in order determine whether Article 22 or s 71 read with s 5 applies to them or not.[51] This problem is made all the more difficult to the extent that it has been recognised, for example, that the wording of Article 22 is capable of leading to inconsistent interpretations and gaps in enforcement.[52]

The Working Party highlights that Article 22 cannot be avoided 'by fabricating human involvement'. These considerations are no less relevant in the context of s 71. In this regard, the Working Party states:

> 'To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than a token gesture'.[53]

It has been pointed out that the assumption underpinning Article 22 is that the pitfalls associated with the use of ADM would be avoided to the extent that there is some measure of human involvement.[54] It has, however, also been highlighted that the regulation of 'solely automated decisions' does not address the 'obvious risks of involving humans in ADM processes, prohibiting the most extreme uses of ADM' while ignoring other 'similarly

[51] Davis, Peter Davis and Sebastian Felix Schwemer 'Rethinking Decisions Under Article 22 of the GDPR: Implications for Semi-Automated Legal Decision-Making' 19 June 2023, available at: *https://ssrn.com/abstract=4478107* or *http://dx.doi.org/10.2139/ssrn.4478107*, accessed on 3 July 2023.
[52] Kaltheuner & Bietti op cit note 41 at 10.
[53] Article 29 Data Protection Working Party op cit note 39 at 21.
[54] Brownlie op cit note 3 at 14.

impactful, uses of the technology'.[55] Crootof, Kaminski & Price[56] make an important point by highlighting that placing a 'human in the loop' simply creates a new type of 'hybrid' system which may not necessarily result in the best of both worlds to the extent that such systems can cause or intensify widely recognized problems that policymakers are either not aware of or which they choose to ignore.[57]

In this context, Brownlie[58] refers to the Wisconsin Supreme Court decision in *State v Loomis,*[59] a case concerned with the COMPAS system mentioned earlier. This system accessed data from a defendant's criminal record together with an interview to generate three 'risk scores'.[60] The judge in Loomis considered various factors, including the COMPAS risk score, when deciding a long prison sentence. The defendant argued that the COMPAS system increased his risk due to gender, constituting 'unconstitutional discrimination on the basis of sex'.[61] The question posed by Brownlie is whether the decision to impose the sentence would fall within the ambit of Article 22 of the GDPR, if such a case had to occur in the EU.[62] She submits that Article 22 (and I would add s 71 in the South African context) would not have applied because the judge did not rely solely on the COMPAS system in arriving at the decision to impose the sentence that he did. Brownlie states that *Loomis* demonstrates the danger in only regulating decisions based solely on automated processing by placing undue emphasis on 'human involvement'.[63] An important issue is that human decision-makers who use 'ADM tools' have no ability to oversee the ADM process and the judge in *Loomis* would not have been in a position to recognize and reduce the potential for bias in the COMPAS system.[64] Importantly, Brownlie makes the point that 'human involvement provided no additional assurance that the COMPAS algorithm was fair and accurate'.[65] Kaltheuner & Bietti also highlight the inherent problem of the black box in the context of human oversight or where a human ostensibly is the ultimate decision-maker.[66] They make the point that 'even if a human being makes the final decision, an automated

[55] Ibid.
[56] Rebecca Crootof, Margot E Kaminski & W Nicholson Price II 'Humans in the Loop' (2023) 76 *Vanderbilt Law Review* 429 at 461.
[57] Ibid.
[58] Ibid.
[59] *State v Loomis* 881 NW 2d 749 (Wis 2016).
[60] Brownlie op cit note 3 at 15.
[61] Ibid.
[62] Ibid.
[63] Ibid.
[64] Ibid at 16.
[65] Ibid.
[66] Kaltheuner & Bietti op cit note 41 at 14.

process has effectively made the decision for them without the human being having the capacity to meaningfully query that decision'.[67]

A further issue identified by Brownlie is that of so-called 'automation bias' where human decision-makers are more likely to accept the automated decision instead of challenging it.[68] These sentiments have been echoed by Castets-Renard [69] who states that in these circumstances it is quite challenging to deviate from a decision arrived at by a machine and then provide a justification for it. Casey, Farhangi & Vogl[70] describe this challenge in the following terms:

> 'Indeed, study after study has documented an innate human tendency to assume the validity of decisions made by algorithms, even when presented with information that directly contradicts the decision's apparent validity.'

Brownlie concludes that because of the restrictive scope of Article 22, its use will be somewhat limited and substantial applications of ADM processes, like the COMPAS algorithm, will escape regulatory control.[71]

Brownlie, may, however, have stated the position somewhat simplistically. Determining where the decision is located (rather than only fixating on whether there is some measure of human involvement) could be more effective in assessing whether Article 22 (or s 71) applies or not.[72] A decision could be located immediately upon the creation of a profile which could give rise to a decision to treat an individual in one way as opposed to another.[73] Binns & Veale[74] refer to the TSA's screening selection test in the US as an example. This process identifies individuals posing potential security risks, leading to possible delays or missed flights. [75] The authors suggest that this initial profiling would have a significant

---

[67] Ibid.
[68] Brownlie op cit note 3 at 16.
[69] Céline Castets-Renard 'Accountability of Algorithms in the GDPR and beyond: A European Legal Framework on Automated Decision-Making' (2019) 30 *Fordham Intellectual Property, Media and Entertainment Law Journal* 91 at 122.
[70] Bryan Casey, Ashkon Farhangi & Roland Vogl 'Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise' (2019) 34 *Berkeley Technology Law Journal* 143 at 146.
[71] Brownlie op cit note 3 at 16.
[72] Reuben Binns & Michael Veale 'Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR' (2021) 11 *International Data Privacy Law* 319 at 326.
[73] Ibid at 325.
[74] Ibid.
[75] Ibid.

effect.[76] Alternatively, the authors suggest that the decision could be located 'downstream', producing a direct effect.[77] For example, where flight risk is 'automatically assigned' but additional screening is subject to the TSA representative's discretion, it is arguable that the decision is made at this subsequent stage. Accordingly, if the initial automatic risk assignment triggers further screening, Article 22 applies whereas if the decision is made at a later stage, following human involvement, Article 22 will not apply.[78]

Almost identical issues have come up for consideration in the matter of *SCHUFA Holding*[79] which is currently pending before the Court of Justice of the European Union (CJEU). The CJEU is expected to provide clarity on ADM in the context of Article 22. SCHUFA, a credit bureau, used mathematical methods to calculate a credit score for a loan applicant, leading to the rejection of her request for credit. The applicant, thereafter, sought details from SCHUFA about how the score was determined. SCHUFA, however, only provided general information about her score without any specific details.

The matter was eventually taken to the Wiesbaden Administrative Court, which stayed proceedings and has asked the CJEU to provide a ruling as to whether Article 22 contemplates that 'the automated establishment of a probability value' in relation to future loan repayments 'constitutes a decision based solely on automated processing, including profiling'.[80] The Administrative Court has argued that credit scoring goes beyond mere profiling and is an independent automated decision,[81] even if a third party makes the final decision.[82] The Court suggested that the credit agency's score heavily influenced the final decision.

The Advocate General[83] has agreed with the Administrative Court and has also downplayed the role of third-party controllers when the provided probability values significantly impact the final decision.

---

[76] Ibid.
[77] Ibid.
[78] Ibid.
[79] Case C-634/21, SCHUFA Holding and Others (Scoring).
[80] Ibid at para II(1).
[81] Ibid at para 21.
[82] Ibid at para 25.
[83] Press Release 49/23 issued by the CJEU on 16 March 2023, available at *https://curia.europa.eu/jcms/p1_3880327/cs/*, accessed on 8 July 2023. The full text of the Advocate General's opinion is not available in English. According to the Press Release, the role of the Advocate General is 'to propose to the Court, in complete independence, a legal solution to the cases for which they are responsible'.

Both the Administrative Court and the Advocate General have clearly adopted a broad interpretation as to what constitutes a 'decision' contemplated in Article 22, downplaying human involvement influenced by automated processing. The CJEU's decision is likely to follow this approach, impacting not only credit agency processing but also ADM in general.

The Working Party, in dealing with the second component of Article 22(1) – the effect of the automated decision, states that a 'legal effect' would be one that affects a person's 'freedom to associate with others, vote, or take legal action' and would also include 'something that affects a person's legal status or their rights under a contract'.[84] The more challenging element is the meaning to be attributed to the phrase 'similarly significantly affects him or her'. In this regard, the Working Party states that even where a person's legal rights or obligations are unchanged by the automated decision, the data subject could still be sufficiently affected to warrant protection under Article 22.[85] In considering the threshold to be applied, the 'significance must be similar to that of a decision producing a legal effect'.[86] Section 71 of the POPIA is somewhat vague and far more imprecise - the decision must affect the data subject to 'a substantial degree'. The Working Party guidance may assist in this regard to the extent that it recognises that what is 'significant' or 'substantial' are not poles apart. In this regard, the Working Party states as follows[87]:

> 'For data processing to significantly affect someone the effects of the processing must be sufficiently great or important to be worthy of attention.'

The Working Party, however, acknowledges that it is hard to be accurate in identifying what would be 'sufficiently significant' to meet the threshold.[88] Some examples, include decisions affecting access to credit (á la *SCHUFA*), access to health services, access to education or decisions denying opportunities to employment.[89] The problem is that it is not clear whether the enquiry is an objective or subjective one and if subjective, whether it's the data subject or the controller that is the focus of the determination.[90] It has been suggested that 'the Working Party has opted for a nuanced subjective interpretation of "significant effects"' and in doing so, 'runs the risk of placing the burden of proof on the data subject'

[84] Article 29 Data Protection Working Party, op cit note 39 at 21.
[85] Ibid.
[86] Ibid.
[87] Ibid.
[88] Ibid at 22.
[89] Ibid.
[90] Kaltheuner & Bietti op cit note 41 at 11.

whereas it should rather 'be for the controllers to ensure that profiling does not significantly affect individuals according to an objective standard'.[91]

It is important to recognize that the GDPR does make some effort to provide a degree of accountability on the part of a controller when a decision is likely to produce a legal effect. In this regard, Article 35(3) provides that a controller must conduct a Data Protection Impact Assessment (DPIA) in circumstances where, inter alia, there will be a 'systematic and extensive evaluation' of personal information which 'is based on automated processing, including profiling' which will 'produce legal effects concerning the natural person or similarly significantly affect the natural person'. The Working Party suggests that a DPIA could enable a controller to ascertain the measures it will adopt to address the 'data protection risks' involved in the relevant processing.[92] Some of these might involve informing the data subject about 'the existence of and the logic involved' in the ADM process, explaining 'the significance and envisaged consequences of the processing', enabling the data subject to 'oppose the decision' and permitting the data subject to 'express their point of view'.[93] None of these possibilities are, however, expressly provided for in Article 35 and the Working Party may, therefore, be somewhat optimistic in assuming that controllers will go to these lengths.

The POPIA does not make provision for a DPIA. Instead, s 58 of the POPIA requires a responsible party to notify the Information Regulator if any processing will fall into the categories specified in s 58. The Regulator is then required to conduct an investigation as to the lawfulness of the processing and, if satisfied, authorize the processing. Only if prior authorization has been obtained, can the processing take place. Amongst the categories referred to in s 58 is the processing of information 'for the purposes of credit reporting' (again, á la *SCHUFA*). Surprisingly, if the Regulator fails to communicate a decision within the prescribed time limits (four weeks for a standard investigation or 13 weeks for a detailed investigation), the responsible party may 'presume a decision in its favour and continue with its processing'.[94] It is therefore reasonably foreseeable that in instances where the intended processing will culminate in an automated decision based on an opaque algorithm, ss 57 and 58 will offer cold comfort to an affected data subject.

---

[91] Ibid.
[92] Article 29 Data Protection Working Party op cit note 39 at 30.
[93] Ibid.
[94] Section 58(7).

In terms of Article 22(3) of the GDPR:

> 'In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision'.

Article 22(3) envisages that a data subject may wish to exercise the right to challenge a decision of the kind referred to in Article 22(1) but only where the decision was necessary for entering into, or performance of, a contract between the data subject and a data controller or the decision was based on the data subject's explicit consent. This, as the Working Party recognises will only be possible if the data subject can understand how the decision was made and how it was arrived at.[95] This will inevitably give rise to the challenges referred to earlier characterized by systems that use black box algorithms.

The Working Party nonetheless recognises the importance of human intervention and that any review should be undertaken by a person who has the requisite authority and ability to change the decision.[96] In carrying out the review, the person should carefully assess the relevant information as well as any further information supplied by the data subject .[97]

It has been suggested that there is 'a need for greater human decision making in Internet infrastructure' in the first place but only if it is meaningful as stipulated by the Working Party referred to earlier.[98] Where a human decision constitutes mere 'rubber stamping' of an automated decision, human rights will be weakened as a result whereas 'meaningful human discretion can contribute to developing a more effective human rights Internet architecture'.[99] Kaltheuner & Bietti[100] suggest that for human intervention to be meaningful, the intervening human must be 'able to determine whether the profile that informs their decision is accurate, fair, and non-discriminatory'. Brennan-Marquez, Levy & Susser[101] identify different categories of 'human influence' or 'human roles' that range from being

[95] Article 29 Data Protection Working Party op cit note 39 at 27.
[96] Ibid.
[97] Ibid.
[98] Ben Wagner 'Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems' (2019) 11 *Policy & Internet* 104 at 118.
[99] Ibid.
[100] Kaltheuner & Bietti op cit note 41 at 14.
[101] Kiel Brennan-Marquez, Karen Levy & Daniel Susser 'Strange Loops: Apparent versus Actual Human Involvement in Automated Decision Making' (2019) 34 *Berkely Technology Law Journal* 745 at 749.

purely ‘procedural’ ‘for example, pushing a given case up or back in the relevant queue, or deciding which cases merit more institutional resources’ to being ‘dispositive’, which would include the ‘power to shape outcomes’. The authors suggest that whilst particular human roles may vary, an important element is that a significant degree of discretion is exercized by a human.[102] The cases referred to in this context are those that have an ‘concrete’ impact on ‘a specific affected party’.[103]

Aside from the challenge presented by the black box, the problem is that in many cases human oversight is not always conducted by suitably qualified individuals. Wagner[104] points out that companies that seek to include a ‘human in the loop’[105] do not, at the same time, ensure that adequately trained staff are able to make these decisions or that they have enough time to do so on their own. In addressing these issues, there should be ‘minimum standards’ which will ensure that simply having a ‘human in the loop’ will no longer suffice as a means to ensuring proper and meaningful oversight.[106]

Section 71(3)[107] provides that:

> ‘the appropriate measures referred to in subsection (2)(a)(ii) must -
>
> (a) provide an opportunity for a data subject to make representations about a decision referred to in subsection (1); and
> (b) require a responsible party to provide a data subject with sufficient information about the underlying logic of the automated processing of the information relating to him or her to enable him or her to make representations in terms of paragraph (a)’.

Section 71(3) goes further than Article 22(3) by requiring that enough information be furnished to a data subject to enable him or her to meaningfully contest a decision. A data subject is also entitled to be furnished with ‘meaningful information about the logic involved’ in an automated decision under the GDPR pursuant to Articles 13, 14 and 15 which is addressed in more detail below, but that entitlement is present irrespective of whether the data subject contests the decision or not. It is likely, however, that a data subject

---

[102] Ibid.
[103] Ibid at 750.
[104] Wagner op cit note 98 at 114.
[105] Ibid.
[106] Ibid at 117.
[107] The ‘appropriate measures’ referred to in s 71(3) refer to the measures to protect a data subject’s legitimate interests where the decision has been taken in connection with the conclusion or execution of a contract.

would want to exercise his or her rights to be furnished with 'meaningful information about the logic involved' in order to give effect to Article 22(3). This, however, is not so clear-cut given the uncertainty surrounding the nature and extent of the information envisaged by Articles 13 to 15 of the GDPR.

It has been suggested that the entitlement to meaningful information envisaged by Articles 13 to 15 of the GDPR only operates ex ante and does not provide a data subject with the entitlement to an ex post facto explanation.[108] Under s 71(3) of the POPIA, the entitlement to the relevant information depends entirely on whether a data subject wishes to make representations concerning an automated decision and therefore clearly envisages the entitlement to an ex post facto explanation. In other words, the right to information about the 'underlying logic of the automated processing', while only available in these limited circumstances, will entitle a data subject to potentially obtain information relating to the rationale of the decision. Again, whether this entitlement would include an explanation of the specific decision is unclear. This important aspect is considered further later.

Article 22 of the GDPR goes much further than s 71 of the POPIA by specifically catering for special categories of personal data referred to in Article 9(1)[109] by providing that automated decision making is only permissible if one of the exemptions in Article 22(2) is applicable and points (a) or (g) of Article 9(2) applies. In this regard, Article 9(2)(a) provides for the explicit consent of the data subject and 9(2)(g) permits, inter alia, the processing of personal data which is 'necessary for reasons of substantial public interest'. In either instance, there have to be suitable measures put in place to protect the data subject's rights and freedoms and legitimate interests. No equivalent provision is found in s 71 the POPIA itself, however, the processing of 'special personal information', in general, is catered for separately in ss 26 to 33 thereof.

---

[108] Sandra Wachter, Brent Mittelstadt & Luciano Floridi 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 76.

[109] Article 9(1) provides that 'Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.'

The difficulties and uncertainties linked to the implementation of Article 22 and s 71(2) are already quite daunting. Yet, these challenges are exacerbated further when combined with the added ambiguities concerning the so-called 'right to an explanation'.

VI. THE RIGHT TO AN EXPLANATION

Article 22, Articles 13 to 15 of the GDPR and s 71(3) of the POPIA are important in relation to a data subject's 'right to an explanation' in the context of ADM. To recap for a moment — Article 22 itself contains no provision catering for the provision of meaningful information in relation how the decision was arrived at. This, some argue, is deliberate and 'suggests that legislators did not intend to implement a right to explanation of specific decisions in the GDPR.'[110] Articles 13 and 14 of the GDPR have been described as specifying 'notification duties for data controllers'.[111] These duties, it has been argued, only operate before a decision has been made on the basis that, by its very nature, a notification takes place prior to a decision when the information is collected.[112] Moreover, the language used in Articles 13(2)(f) and 14(2)(g) to the effect that the data subject must be informed about the existence of ADM and be provided with 'meaningful information about the logic involved as well as the significance and the envisaged consequences of such processing' implies that such information is confined to general information regarding the functionality of the automated decision-making system, its purpose and its anticipated effect.[113]

Article 15 has been described as providing a data subject with a right of access, which can be invoked by a data subject at any time which would include any point after a decision has been made. This would imply that, in contrast to Articles 13 and 14, a data subject would be entitled to an ex post facto explanation of a specific decision that has already been arrived at.[114] It has, however, been argued that:

> '[t]he phrase "envisaged consequences" is future oriented, suggesting that the data controller must inform the data subject of possible consequences of the automated decision-making before such processing occurs. This interpretation follows the timeline constraints

[110] Wachter, Mittelstadt & Floridi op cit note 108 at 81.
[111] Ibid at 82.
[112] Ibid.
[113] Ibid.
[114] Ibid at 83.

of identical provisions in Articles 13(2)f and 14(2)g discussed above, which only allow for ex ante explanations.'[115]

It has also been suggested that the information to which a data subject is entitled will be confined to an explanation as to how the system functions.[116]

Unfortunately, the Working Party Guidelines do not provide a definitive answer one way or another. It has been argued, however, that the Guidelines (albeit when they were in draft form),[117] 'interpret "meaningful information about the logic involved" as an ex ante right about system functionality, for both, Articles 13(2)(f) and 14(2)(g), as well as Article 15(1)(h).'[118]

An ex ante entitlement to information 'about the logic involved' may be easier to digest in the context of a data subject's notification rights to the extent that these rights relate to processing that will take place in the future. It is also relatively palatable to similarly accept that the nature of the information will, in general terms, be limited to system functionality given that no decision at that point in time will have been made. It is, however, difficult to understand the basis upon which a data subject's access rights which can be exercised after a decision has been made should not include an ex post facto explanation of the specific decision. A general recital of how the system works in these circumstances is unlikely to enable a data subject to properly contest a decision as provided for in Article 22(3) of the GDPR or s 71(3)(b) of the POPIA.

Some, however, have argued that the vague and broad language used in Articles 13 to 15 and Article 22 may favour a wider interpretation of these provisions over a restrictive interpretation which would seriously impact data controllers.[119] The same scholars, however, recognise that Articles 13 and 14 of the GDPR only operate before the processing of personal data and can only result in general information being disclosed whereas Article 15 might result in the disclosure of information pertaining to a specific decision in relation to a data subject.[120] They also argue that Article 15 envisages the furnishing by a data

---

[115] Ibid.
[116] Ibid at 84.
[117] The draft Guidelines are substantively identical to the current version.
[118] Kaltheuner & Bietti op cit note 41 at 15.
[119] Gianclaudio Malgieri & Giovanni Comandé 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 243 at 247.
[120] Ibid at 255.

controller of information 'about the logic actually employed, and not only about the general system functionality of an algorithmic decision-making'.[121] On the other hand, Selbst & Barocas[122] state the reference in Articles 13 to 15 to the 'logic' of the decision, only envisages an explanation that focuses on the general functionality of a model, as opposed to how it arrives at a specific result.

There is clearly enough ambiguity to support both points of view. Although Wachter, Mittelstadt & Floridi[123] vigorously argue that the right to an explanation is limited and that a meaningful right to an explanation of 'specific automated decisions' is not provided for in the GDPR, they state that considerations such as accountability and transparency 'may provide compelling reasons for legislators or data controllers to introduce one in the future'.[124] The most plausible instance where this could occur would be in the context of Article 22(3) so as to enable a data subject to meaningfully contest or express a view in relation to the rationale of an automated decision.[125] Another plausible ('but unlikely') instance is that in the 'future' the right of access embodied in Article 15 is interpreted broadly to include some aspects of the reasoning or context of a particular decision.[126]

The CJEU in *SCHUFA* may, however, provide much needed clarity much sooner. The Administrative Court, as part of its referral submissions has contended that a data subject is entitled to an ex post facto explanation of a decision pertaining to his or her credit score pursuant to Article 15 to the extent that the automated processing of such score falls within the ambit of Article 22.[127] The Advocate General has also weighed in on this aspect and has submitted that the obligation to provide 'meaningful information about the logic involved' must be adequately detailed in relation to the methodology employed and must include 'general' information on the factors and their 'aggregate' weighting that were taken into account in relation to the specific decision to enable the individual to challenge it. [128]

The distinction between the adoption of a narrow approach regarding the extent of a right to an explanation as opposed to a more expansive interpretation, could potentially be crucial

---

[121] Ibid at 256.
[122] Andrew D Selbst & Solon Barocas 'The Intuitive Appeal of Explainable Machines' (2019) 87 *Fordham Law Review* 1085 at 1107.
[123] Wachter, Mittelstadt & Floridi op cit note 108.
[124] Ibid at 90.
[125] Ibid at 91.
[126] Ibid.
[127] SHUFA op cit note 79 at paras 29 to 32.
[128] Press Release 49/23 op cit note 83.

in situations where the relevant decision is incapable of being explained. This issue has spawned significant debate but yielded very little in terms of definitive answers.

VII. THE CHALLENGE

Even if a broader and more expansive interpretation is afforded to Articles 15 and 22 (and s 71 of the POPIA), a far more challenging issue is whether the rationale of the specific automated decision is capable of being explained given that most machine learning models upon which ADM relies are opaque as alluded to earlier.[129] In fact, the higher the degree of automation, the likelihood of opacity will increase which makes it far more challenging to provide explanations to humans which undermines transparency.[130] Even though the CJEU is likely to rule that an individual will be entitled to information that relates to the specific decision, for as long as ADM systems are based on complicated and impenetrable algorithms, these requirements may be difficult, if not, impossible to implement.

Whilst these issues are still somewhat in flux but have nonetheless, spawned significant debate and analysis in the context of the GDPR, the challenges in relation to a data subject's right to a meaningful explanation have yet to be considered specifically in the context of s 71 of the POPIA. At this early stage of the POPIA's development, Neethling, Potgieter & Roos[131] and Burns & Burger-Smidt[132] have referenced s 71 in somewhat neutral terms and appear to have adopted a 'wait and see' stance before weighing in on this issue. Others have highlighted the important implications that AI holds in the context of the South African Constitution such as, inter alia, transparency, accountability, privacy and non-discrimination.[133]

Bronstein[134] has, however, specifically highlighted that ADM may give rise to challenges in relation to the remedies available to individuals. Bronstein states that both Article 22 and s 71 'allow for some sensible dialogue with the data controller or responsible party' but that

---

[129] Rudin & Radin op cit note 35 at 3.
[130] Tal Zarsky 'The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making' (2016) 41 *Science, Technology, & Human Values* 118 at 121.
[131] J Neetling, JM Potgieter & A Roos *Neetling on Personality Rights* at 398.
[132] Yvonne Burns & Ahmore Burger-Smidt *A Commentary on the Protection of Personal Information Act* at 186.
[133] Dirk J Brand 'Artificial Intelligence in Government: Development of a Legal Framework for South Africa' (2022) 14 *eJournal of eDemocracy and Open Government* at 142-145.
[134] Victoria Bronstein 'Prioritising Command-and-Control Over Collaborative Governance: The Role of the Information Regulator Under the Protection of Personal Information Act (2022) 25 *PELJ* 1.

these remedies 'are likely to be disappointing'.[135] The problem identified by Bronstein is the inadequacy of so-called 'explainable artificial intelligence' and the 'reliance on transparency' and the fact that these may place an excessive burden on 'users' who are unlikely to have the power to change the outcome.[136] The difficulty is that most AI models that underpin ADM are based on opaque algorithms and that therefore very little 'sensible' or more importantly, meaningful 'dialogue' as suggested by Bronstein is, in fact, possible. At the moment, 'explainable AI' (XAI), for the most part, is still aspirational and so the challenge has been misidentified.[137] As part of addressing the right to an explanation as embodied in the GDPR, XAI has, in fact, been postulated as a solution to address the lack of transparency in most AI models. [138] In this context, the GDPR has been identified as an important driver for XAI research.[139] Whilst it may be true that even in the context of XAI, an individual may be disempowered to challenge a decision based on ADM, that potential difficulty highlights a different problem, perhaps at a broader societal level. If XAI is achieved, and a meaningful explanation is capable of being provided to an individual in relation to an automated decision, an outcome which is not so clear-cut, the question of whether he or she ultimately succeeds in contesting the decision is a different topic for a different discussion.

VIII. THE WAY FORWARD

As mentioned, a possible solution to the challenge presented by black box models is XAI, the goal of which is to create 'transparent and clear decision-making processes and decisions, which in turn should result in trust in decisions'.[140] There are some who have gone so far as to suggest that black box models are deliberately used even when a simpler and far more understandable model is capable of being used without sacrificing accuracy and reliability.[141] Rubenstein[142] is less optimistic that XAI is a viable solution, calling it 'faux techno-solutionism' which may have some benefit to data scientists but will offer very

---

[135] Ibid at 19.
[136] Ibid.
[137] Hans de Bruijn, Martijn Warnier & Marijn Janssen 'The Perils and Pitfalls of Explainable AI: Strategies for Explaining Algorithmic decision-making' (2022) 39 *Government Information Quarterly* 1 at 2.
[138] Ashley Deeks 'The Judicial Demand for Explainable Artificial Intelligence' (2019) 119 *Columbia Law Review* 1829 at 1829.
[139] Maja Brkan & Grégory Bonnet 'Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas' (2020) 11 *European Journal of Risk Regulation* 18 at 19.
[140] Ibid at 4.
[141] Rudin & Radin op cit note 35 at 3.
[142] David S Rubenstein 'Acquiring Ethical AI' (2021) 73 *Florida Law Review* 747 at 791.

little insight to 'end users, auditors, adjudicators, and policymakers'. Whether XAI is a complete or partial solution in the context of providing a meaningful or sufficient explanation to a data subject is very much an open question, an issue that falls outside the ambit of this analysis. In fact, even the meaning of the term 'explanation' is not universally accepted. It could encompass an 'interpretation' as to how a decision was arrived whereas a 'justification' would focus on why a decision arrived at was 'correct' or 'reasonable'.[143]

Bronstein suggests the that requirement to carry out a DPIA under the GDPR in the case of so-called 'high-risk' processing would require 'technical staff' within organizations to 'describe and explain their systems'.[144] This, she postulates would be conducive to 'more informed decision-making' and demonstrate that controllers have acted with 'due diligence'.[145] Bronstein believes that to the extent a DPIA is 'executed in a transparent way' it can help build confidence and demonstrate an organization's compliance with all its obligations from a legal, ethical and human rights perspective.[146] It would seem this view is informed by the idea that these auditing mechanisms would enable the development and use of algorithmic systems that are transparent and fair.[147] The obvious question is the extent to which this supposed outcome could stifle innovation. But, in any event, a DPIA is only required to include 'a systematic description of the envisaged processing operations' which does not necessarily imply that the use of so-called black box models will suddenly come to an end. The requirement to carry out a DPIA is imposed on a controller and takes place outside the ambit of Article 22 read with Articles 13 to 15 and therefore does not affect a data subject's independent right to a meaningful explanation. It is therefore unlikely that the requirement to carry out a DPIA will address or solve the challenges associated with giving effect to this right. Certainly, from a South African perspective, this debate will, in any event, be of academic interest only.

Given the lack of clarity on what a meaningful explanation encompasses and the fact that it could be limited to information on system functionality,[148] which would be of little use to a data subject, the focus should perhaps move to what would benefit a data subject the most in the context of ADM. It has been suggested that the most beneficial explanation would

[143] Brkan & Bonnet op cit note 139 at 25 to 26.
[144] Bronstein op cit note 134 at 21.
[145] Ibid.
[146] Ibid.
[147] Casey, Farhangi & Vogl op cit note 70 at 183.
[148] Wachter, Mittelstadt & Floridi op cit note 123 at 81.

be one that assists a data subject in understanding why a decision was made, provides 'grounds' to contest a decision and helps in understanding what could be done to change a decision going forward.[149] Explanations of this type would not require an explanation of 'the internal logic of automated decision-making systems'.[150] The idea is that providing a data subject with 'counterfactuals' ought to be the way in which explanations are given.[151] A simple example of a counterfactual explanation expressed by Wachter, Mittelstadt & Russell essentially encapsulates it thus — 'You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan.'[152] This, they say will avoid having to explain 'the internal workings of complex machine learning systems' and will provide information to a data subject which is 'digestible' and 'practically useful for understanding the reasons for a decision'.[153] The authors suggest that counterfactuals will facilitate compliance with the GDPR without having to delve into issues of 'interpretability' or having to grapple with black box systems.[154]

The provision of a counterfactual is certainly attractive to the extent that it avoids many of the pitfalls addressed in this analysis. Since the GDPR and the POPIA offer very little guidance on precisely what type of explanation should be provided to a data subject, there is definitely room to manoeuvre if one accepts that the GDPR and the POPIA's purpose is to protect the rights of individuals. Giving effect to an individual's right to an explanation will be better served if the explanation is capable of assisting a data subject who has been adversely affected by ADM. This will not be achieved by reciting how, for example, the algorithm works and the variables and weightings it took into account, if this information is not understandable. Even XAI may prove to have little relevance or value to an individual if all it does is explain the internal workings of a system more simply. From a practical perspective, the advantage of providing an individual with a counterfactual is that it breaks down, in simplistic terms, exactly why a decision was made and how the outcome could have been different. Of course, this could uncover inherent systemic issues, if the

---

[149] Sandra Wachter, Brent Mittelstadt & Chris Russell 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31 *Harvard Journal of Law & Technology* 841 at 843
[150] Ibid.
[151] Ibid at 844.
[152] Ibid.
[153] Ibid at 860.
[154] Ibid at 861.

counterfactual exposes bias or discrimination, which is a subject in its own right, but at least these aberrations would have been exposed more easily and transparently.

IX. CONCLUSION

The GDPR and the POPIA, in seeking to protect data subjects from being subject to ADM without any recourse in relation to decisions which have legal effect, or which are significant or substantial, do not clearly prescribe the nature and extent of what comprises the so-called right to an explanation. To the extent that this right is limited to basic information concerning the mechanics of the algorithm, it is doubtful that a data subject would be able to meaningfully contest an automated decision. To the extent that a broad interpretation is given to Articles 15 and 22 of the GDPR and s 71(3) of the POPIA, the use of black box AI, may preclude the right to a meaningful or sufficient explanation. In both instances, having a human in the loop takes the matter no further. Perhaps, the emphasis should rather focus on providing a data subject with a counterfactual which will provide a data subject with an explanation that is truly 'meaningful' and capable of having practical value. In this way, giving effect to the right to an explanation in this manner will better serve data subjects without stifling innovation and the development of AI systems. There is no doubt that both the GDPR and the POPIA, whilst attempting to regulate and protect data subjects from the adverse effects of ADM, have not gone far enough by failing to adequately prescribe the type of information a data subject can expect to receive when invoking the right to an explanation. At least s 71 of the POPIA, unlike Article 15 of the GDPR, envisages that the information furnished to a data subject about the underlying logic of the decision, must be sufficient to enable him or her to make representations in relation to the decision. The vagueness and impreciseness, however, still persists in both provisions as to what is meant by the 'underlying logic' pertaining to a decision. Given the continual debate on this issue, there is clearly a need for this lacuna to be addressed by subsequent iterations of the respective pieces of legislation. Until then, it is hoped that the courts, such as the CJEU in *SCHUFA*, will creatively address these challenges head-on and provide a measure much-needed certainty in the interim.

Word Count: 10 704

BIBLIOGRAPHY

Books

Burns Y & Burger-Smidt A *A Commentary on the Protection of Personal Information Act* (2018) LexisNexis

Neetling J, Potgieter JM & Roos A *Neetling on Personality Rights* (2019) LexisNexis

Russell S J & Norvig P *Artificial Intelligence A Modern Approach* 3 ed (2009) Prentice Hall

Chapters in edited collections

Bosco F, Creemers N & Farraris V et al 'Profiling Technologies and Fundamental Rights and Values: Regulatory Challenges and Perspectives from European Data Protection Authorities' in Gutwith S, Leenes R & de Hert P et al (eds) (2015) *Reforming European Data Protection Law. Law, Governance and Technology Series* 3.

Journal articles

Allen R & Masters D 'Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making' (2020) 20 *ERA Forum* 585.

Asay C D 'Artificial Stupidity' (2020) 61 *William & Mary Law Review* 1187.

Binns R 'Human Judgment in algorithmic loops: Individual justice and automated decision-making' (2022) 16 *Regulation & Governance* 197.

Binns R & Veale M 'Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR' (2021) 11 *International Data Privacy Law* 319.

Brand D J 'Artificial Intelligence in Government: Development of a Legal Framework for South Africa' (2022) 14 *eJournal of eDemocracy and Open Government* 130.

Brennan-Marquez K, Levy K & Susser D 'Strange Loops: Apparent versus Actual Human Involvement in Automated Decision Making' (2019) 34 *Berkely Technology Law Journal* 745.

Brkan M & Bonnet G 'Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas' (2029) 11 *European Journal of Risk Regulation* 18.

Bronstein V 'Prioritising Command-and-Control Over Collaborative Governance: The Role of the Information Regulator Under the Protection of Personal Information Act (2022) 25 *PELJ* 1.

Brownlie E 'Encoding Inequality: The Case for Greater Regulation of Artificial Intelligence and Automated Decision-Making in New Zealand' (2020) 51 *Victoria University of Wellington Law Review* 1.

Brynjolfsson E & McAfee A 'The Business of Artificial Intelligence' (2021) *Harvard Business Review Special Issue* 20.

Casey B, Farhangi A & Vogl R ‘Rethinking Explainable Machines: The GDPR’s “Right to Explanation” Debate and the Rise of Algorithmic Audits in Enterprise’ (2019) 34 *Berkeley Technology Law Journal* 143.

Castets-Renard C ‘Accountability of Algorithms in the GDPR and beyond: A European Legal Framework on Automated Decision-Making’ (2019) 30 *Fordham Intellectual Property, Media and Entertainment Law Journal* 91.

Crootof R, Kaminski M E & Price II W N ‘Humans in the Loop’ (2023) 76 *Vanderbilt Law Review* 429.

Dalkilic E E ‘A Way Out For Europe: How can Europe Combat Discrimination by Automated Decision Systems’ (2022) 9 *Lincoln Memorial University Law Review* 37.

Davis P & Schwemer S F ‘Rethinking Decisions Under Article 22 of the GDPR: Implications for Semi-Automated Legal Decision-Making’ 19 June 2023, available at: *https://ssrn.com/abstract=4478107 or http://dx.doi.org/10.2139/ssrn.4478107*.

de Bruijn H, Warnier M & Janssen M ‘The Perils and Pitfalls of Explainable AI: Strategies for Explaining Algorithmic decision-making’ (2022) 39 *Government Information Quarterly* 1 at 2.

Deeks A ‘The Judicial Demand for Explainable Artificial Intelligence’ (2019) 119 *Columbia Law Review* 1829.

Ghosh A, Chakraborty D & Law A ‘Artificial intelligence in Internet of things’ (2018) 3 *CAAI Transactions on Intelligence Technology* 208.

Goertzel B ‘Artificial General Intelligence: Concept, State of the Art, and Future Prospects’ (2014) *Journal of Artificial General Intelligence* 5(1) 1.

Gravett W ‘Bias and lack of accuracy in risk-assessment software in the United States criminal justice system’ (2021) 34 *South African Journal of Criminal Justice* 31.

Huq A Z ‘Constitutional Rights in the Machine-Learning State’ (2020) 105 *Cornell Law Review* 1875.

Jackson M C ‘Artificial Intelligence & Algorithmic Bias: The Issues With Technology Reflecting History & Humans’ (2021) 16 *J.Bus & Tech* 299.

Johnson K, Pasquale F & Chapman J ‘Artificial Intelligence, Machine Learning and Bias in Finance: Toward Responsible Innovation’ (2019) 88 *Fordham Law Review* 499.

Kaltheuner F & Bietti E ‘Data is power: Towards additional guidance on profiling and automated decision-making in the GDPR’ (2018) 2 *Journal of Information Rights, Policy and Practice* 1.

Kroll J A, Barocas S, Felton E W, Reidenberg J R, Robinson D G & Yu H ‘Accountable Algorithms’ (2017) 165 *University of Pennsylvania Law Review* 633.

Lu S ‘Data Privacy, Human Rights and Algorithmic Opacity’ (2022) 110 *California Law Review* 2087.

Malgieri G & Comandé G ‘Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation’ (2017) 7 *International Data Privacy Law* 243.

Palace V M 'What if Artificial Intelligence Wrote This: Artificial Intelligence and Copyright Law' (2019) 71 *Florida Law Review* 217.

Ribeiro M T, Singh S & Guestrin C 'Why Should I Trust You?': Explaining the Predictions of Any Classifier' *arXiv* (2016), available at: *http://arxiv.org/abs/1602.04938*.

Rich M L 'Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment' (2016) 164 *University of Pennsylvania Law Review* 871.

Rubenstein D S 'Acquiring Ethical AI' (2021) 73 *Florida Law Review* 747.

Rudin C & Radin J 'Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson from an Explainable AI Competition' (2019) 1 *Harvard Data Science Review* 3.

Selbst A D & Barocas S 'The Intuitive Appeal of Explainable Machines' (2019) 87 *Fordham Law Review* 1085.

Shackelford S J & Dockery R 'Governing AI' (2020) 30 *Cornell Journal of Law & Public Policy* 279.

Silveira A 'Automated individual decision-making and profiling [on Case C-634/21 – SCHUFA (SCHUFA (Scoring)]' (2023) 8 *EU Law Journal* 74.

Sonia K Katyal S K 'Private Accountability in the Age of Artificial Intelligence' (2019) 66 *UCLA Law Review* 54.

Surden H 'Machine Learning and Law' (2014) 89 *Washington Law Review* 87.

Surden H 'Artificial Intelligence and Law: An Overview' (2019) 35 *Georgia State University Law Review* 1305.

Tschider C A 'Beyond the "Black Box"' (2021) 98 *Denver Law Review* 683.

Wachter B, Mittelstadt B & Russell C 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31 *Harvard Journal of Law & Technology* 841.

Wachter S, Mittelstadt B & Floridi L 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 76.

Wagner B 'Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems' (2019) 11 *Policy & Internet* 104.

Waldman A E 'Power and Automated Decision-Making' (2019) 88 *Fordham Law Review* 613.

Washington A L 'How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica debate' (2019) 17 *Colorado Technology Law Journal* 1.

Whittlestone J, Arulkumaran K & Crosby M 'The Societal Implications of Deep Reinforcement Learning' (2021) 70 *Journal of Artificial Intelligence Research* 1003.

Wiedemann K 'Profiling and (automated) decision-making under the GDPR: A two-step approach' (2022) 45 *Computer Law & Security Review* 1.

Yu P K ‘The Algorithmic Divide and Equality in the Age of Artificial Intelligence’ (2020) 72 *Florida Law Review* 331.

Zarsky T ‘The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making’ (2016) 41 *Science, Technology, & Human Values* 118.

Foreign case law

State v Loomis 881 NW 2d 749 (Wis 2016).

Case C-634/21, *SCHUFA Holding and Others (Scoring)*, available at: *https://curia.europa.eu/juris/showPdf.jsf?text=&docid=250522&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=841423*.

South African Legislation

Protection of Personal Information Act 4 of 2013.

Foreign Legislation

General Data Protection Regulation (EU) 2016/679.

Foreign Guidelines

‘The Privacy Expert’s Guide to AI and Machine Learning’ *Future of Privacy Forum* October 2018 available at *https://iapp.org/resources/article/the-privacy-experts-guide-to-ai-and-machine-learning*.

Article 29 Data Protection Working Party ‘Guidelines on Automated Individual Decision Making and Profiling for the Purpose of Regulation 2016/679’, 6 February 2018, available at: *https://ec.europa.eu/newsroom/article29/items/612053*.