McGill Law Journal — Revue de droit de McGill

A Reasonable Apprehension of AI Bias: Lessons from *R. v. R.D.S.*

*Teresa Scassa**

In 1997, the Supreme Court of Canada rendered a split decision in a case in which the central issue was whether an African Nova Scotian judge who had brought her lived experience to bear in a decision involving a confrontation between a black youth and a white police officer had demonstrated a reasonable apprehension of bias. The case, with its multiple opinions across three courts, teaches us that identifying bias in decision-making is a complex and often fraught exercise.

Automated decision systems are poised to dramatically increase in use across a broad range of contexts. They have already been deployed in immigration and refugee determination, benefits allocation, and in assessing recidivism risk. There is also a growing use of AI-assistance in human decision-making that deserves scrutiny. For example, generative AI systems may introduce dangerous unknowns when it comes to the source and quality of briefing materials that are generated to inform decision-makers. Bias and discrimination have been identified as key issues in automated decision-making, and various solutions have been proposed to prevent, monitor, and correct potential issues of bias. This paper uses *R. v. R.D.S.* as a starting point to consider the issues of bias and discrimination in automated decision-making processes, and to evaluate whether the measures proposed to address bias and discrimination are likely to be effective. The fact that *R. v. R.D.S.* does not come from a decisional context in which we currently use AI does not mean that it cannot teach us—not just about bias itself—but perhaps more importantly about how we think about and process issues of bias.

En 1997, la Cour suprême du Canada a rendu une décision partagée dans un dossier où la question principale était de déterminer si une juge afro-néo-écossaise avait fait preuve d'une crainte raisonnable de biais en valorisant son expérience vécue pour rendre une décision. L'affaire dans laquelle la juge tranchait concernait une confrontation entre un jeune noir et un policier blanc. Ce dossier, et ses plusieurs opinions à travers trois tribunaux, nous enseigne que l'identification du biais dans la prise de décision est une tâche complexe et souvent difficile.

Les systèmes de décision automatisés — qu'ils soient entièrement automatisés ou assistés par l'intelligence artificielle (IA) — sont placés à être utilisés d'avantage à travers plusieurs contextes. Des systèmes de décision automatisés ont déjà été déployés dans les domaines de l'immigration et de la détermination du statut de réfugié, de l'attribution des bénéfices et de l'évaluation du risque de récidivisme. L'utilisation croissante d'assistance de l'IA dans la prise de décision humaine mérite également d'être examinée de près. Par exemple, l'origine et la qualité des rapports produits par des systèmes d'IA générative qu'utilisent les décideurs peuvent introduire des erreurs dangereuses. Le biais et la discrimination ont été identifiées comme des problèmes clés dans la prise de décision automatisée, et plusieurs solutions ont été proposées pour éviter, surveiller et corriger des potentiels problèmes de biais. Cet article utilise l'affaire *R. c. R.D.S.* comme point de départ pour examiner les questions de biais et de discrimination dans les processus de prise de décision automatisée et pour évaluer si les mesures proposées sont susceptibles d'être efficaces. Le fait que l'affaire *R. c. R.D.S.* n'est pas issue d'un contexte décisionnel dans lequel nous utilisons l'IA ne signifie pas qu'elle ne peut pas nous apprendre - non seulement sur le biais — mais peut-être plus importamment, comment nous envisageons et traitons les questions de biais.

* Canada Research Chair in Information Law and Policy, University of Ottawa. I am grateful for helpful feedback on an earlier draft of this paper from Jennifer Raso, as well as for comments on a more recent version from participants at the McGill Law Journal 2024 Symposium, and from the anonymous peer reviewers. Many thanks to the excellent editorial team at the McGill Law Journal. This paper was written without the use of artificial intelligence. I gratefully acknowledge the support of the Canada Research Chairs Program.

Citation: (2024) 69:4 McGill LJ 467 — Référence : (2024) 69:4 RD McGill 467

Introduction

The risk of discriminatory bias is a central concern when it comes to the use of artificial intelligence (AI) technologies for automated decision-making (ADM).[1] Identification and mitigation of such bias is an important preoccupation of emerging laws and policies. Currently, public-sector automated decision systems (ADS) operate in lower risk contexts than the criminal justice system, although the use of such tools is evolving.[2] In the private sector, ADS are already deployed in higher impact contexts such as the selection of tenants for apartments,[3] the determination of creditworthiness,[4] and in hiring and performance evaluation.[5] Generative AI systems such as ChatGPT can also be used to support ADM in different ways, including in the preparation of briefing materials, translation, and drafting decisions.[6]

---

1 See e.g. Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St Martin's Press, 2018) at 6–7, 11–13; Yarden Katz, *Artificial Whiteness: Politics and Ideology in Artificial Intelligence*, (New York: Columbia University Press, 2020) at 8–11; Frederik Zuiderveen Borgesius, *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making* (Strasbourg: Council of Europe, Directorate General of Democracy, 2018) at 15–23; Hugo Cossette-Lefebvre & Jocelyn Maclure, "AI's Fairness Problem: Understanding Wrongful Discrimination in the Context of Automated Decision-Making" (2023) 3:4 AI & Ethics 1255 at 1259–61.

2 For examples of the use of public sector ADM, see Eubanks, *supra* note 1; Jennifer Raso, "Displacement as Regulation: New Regulatory Technologies and Front-Line Decision-Making in Ontario Works" (2017) 32:1 CJLS 75; Law Commission of Ontario, *Regulating AI: Critical Issues and Choices* (Toronto, April 2021).

3 See e.g. "Victoria Startup Creates Artificial Intelligence to Help Landlords Screen Tenants", *CTV News* (9 March 2018), online: <vancouverisland.ctvnews.ca> [perma.cc/S5J6-26LZ]; Rentify, Business Wire News Release, "Rentify Launches AI Tool to Empower Property Managers to Screen Prospective Tenants Using Bank Data" *Financial Post* (13 April 2021), online: <financialpost.com> [perma.cc/U6Y7-MWYR].

4 Marc Schmitt & Marc Roper, "Artificial Intelligence (AI)-Enabled Credit Scoring in Banking and Fintech: In Search of Maximum Prediction Accuracy" (25 July 2023), online: <papers.ssrn.com> [perma.cc/KG9Q-6ZQY].

5 See e.g. Ben Dattner et al, "The Legal and Ethical Implications of Using AI in Hiring", *Harvard Business Review* (25 April 2019), online: <hbr.org> [perma.cc/23WM-8VGQ]; Tiago Jacob Fernandes França et al, "Artificial Intelligence Applied to Potential Assessment and Talent Identification in an Organisational Context" (2023) 9 Heliyon 1 at 3, 20.

6 See e.g. US, California Government Operations Agency, *Benefits and Risks of Generative Artificial Intelligence Report* (November 2023) at 11–12; Treasury Board of Canada Secretariat, "Guide on the use of Generative AI" (21 March 2024), online: <canada.ca> [perma.cc/T8YB-VR8R].

This paper explores discriminatory bias in ADM using the series of court decisions in *R. v. R.D.S.*[7] (*RDS*) (culminating in a 1997 decision of the Supreme Court of Canada) to illustrate some of the potential frailties in approaches to this issue. It is important to note at the outset that *RDS* addressed the issue of 'reasonable apprehension of bias', which differs significantly from the human-rights-based concept of discriminatory bias. Nevertheless, the case is important because in it, the concept of impartiality that underlies the doctrine of reasonable apprehension of bias becomes intertwined with the notion of discriminatory bias in complex and interesting ways. In *RDS*, the alleged apprehension of bias is tied to a Black woman judge's perception of the credibility of two witnesses – one White and one Black. Credibility – something typically stripped from the targets of discrimination – is left to be determined by a decision-maker who is in turn challenged for bringing a racialized (i.e., non-White) perspective to the task. This complicated and messy case challenges risk-mitigation approaches to AI bias in which we identify risks, develop strategies to mitigate them, and monitor outcomes.[8] Risk-based approaches tend to assume that there is a social consensus about what bias is and how it is manifested. They also tend to lead us towards technological solutions. *RDS* teaches us that understanding, identifying, and addressing bias may be much messier.[9]

This paper begins with a brief overview of discriminatory bias in AI systems. Part 2 provides a summary of the dispute at the heart of *RDS*. Part 3 teases out four themes emanating from *RDS* that are relevant to the AI context: (1) the tension between facts and opinion, (2) transparency and explainability, (3) the issue of biased input and biased output and (4) the role of the human-in-the-loop. The paper concludes by considering that a statistical and technological approach to identifying and mitigating bias in ADM may unduly narrow the focus, and argues for a more robust approach to addressing bias in AI.

---

7 *R v RDS*, 1995 CanLII 9321 (NSSC) [*R v RDS* SC]; *R v RDS*, 1995 NSCA 201 [*R v RDS* CA]; *R v S (RD)*, 1997 CanLII 324 (SCC) [*R v RDS* SCC].

8 Margot E Kaminski, "Regulating the Risks of AI" (2023) 103:5 BUL Rev 1347 at 1350–52. Beyond risk mitigation, it is possible to develop AI technologies to specifically counter known biases (see, e.g. Orly Lobel, *The Equality Machine: Harnessing Digital Technology for a Brighter, More Inclusive Future* (New York: PublicAffairs, 2022)).

9 This is a point made by Sujith Xavier in an article that examines the reasonable apprehension of bias test and how it is applied in racialized contexts (see Sujith Xavier, "Biased Impartiality: A Survey of Post-*RDS* Caselaw on Bias, Race and Indigeneity" (2021) 99:2 Can Bar Rev 354).

I. Bias and AI

It is well understood that AI technologies raise problems of harmful or discriminatory bias that must be addressed.[10] The US National Institute of Standards and Technology (NIST) Framework for AI Risk Management identifies "harmful bias" as an issue of fairness, linking it to concerns for equality and equity.[11] It identifies three broad categories of bias in AI: "systemic, computational and statistical, and human-cognitive,"[12] all of which can be present without any intention to discriminate. *Systemic bias* can be found "in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems."[13] *Statistical or computational bias* is "anything that leads to a systematic difference between the true parameters of a population and the statistics used to estimate those parameters."[14] Errors that create statistical bias can arise from the collection of the data, its classification, the omission of certain variables, or choices made in study design or in the weighting of different variables. *Human cognitive biases*, according to the NIST Framework:

> relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about purposes and functions of an AI system. Human-cognitive biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI.[15]

Both systemic and human cognitive bias are more complex than statistical bias, and neither can be eliminated through the correction of datasets or algorithms. For example, in considering systemic bias, the NIST Framework notes that "systems in which predictions are somewhat balanced across demographic groups may still be inaccessible to individuals

---

10 See e.g. *Artificial Intelligence and Data Act*, being Part 3 of the Bill C-27, *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*, 1st Sess, 44th Parl, 2022 (first reading 16 June 2022) [*AIDA*]; National Institute of Standards and Technology US Department of Commerce, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (Maryland: National Institute of Standards and Technology, 2023) [NIST, *AI RMF*].

11 NIST, *AI RMF*, *supra* note 10 at 17.

12 *Ibid* at 18.

13 *Ibid*.

14 Jenny Gutbezhal, "5 Types of Statistical Bias to Avoid in Your Analyses" (13 June 2017), online: <online.hbs.edu> [perma.cc/LSF3-XYKS].

15 NIST, *AI RMF*, *supra* note 10 at 18.

with disabilities or affected by the digital divide or may exacerbate existing disparities or systemic biases."[16]

A recent initiative between the EU and the US to harmonize AI terminology addresses both harmful bias and discrimination, and distinguishes between the two:

> Harmful AI bias describes systematic and repeatable errors in AI systems that create unfair outcomes, such as placing privileged groups at systematic advantage and unprivileged groups at systematic disadvantage. Different types of bias can emerge and interact due to many factors, including but not limited to, human or system decisions and processes across the AI lifecycle. Bias can be present in AI systems resulting from pre-existing cultural, social, or institutional expectations; because of technical limitations of their design; by being used in unanticipated contexts; or by non-representative design specifications.[17]

The sources of bias are disparate, making it challenging to address. The EU-US definitions identify discrimination as a subset of harmful bias. Discrimination is defined as a form of unequal treatment that "can be a result of societal, institutional and implicitly held individual biases or attitudes that get captured in processes across the AI lifecycle, including by AI actors and organisations, or represented in the data underlying AI systems."[18] The definition of discrimination goes on to note that discriminatory bias

> can also emerge due to technical limitations in hardware or software, or the use of an AI system that, due to its context of application, does not treat all groups equally. Discriminatory biases can also emerge in the very context in which the AI system is used. As many forms of biases are systemic and implicit, *they are not easily controlled or mitigated and require specific governance and other similar approaches*.[19]

The essential difference between harmful bias and discrimination appears to be agency. Discrimination comes from human biases that infect AI processes or design, whereas harmful bias flows from flawed data or design issues, some of which may be impacted by the social context in which they are developed (which in turn suggests at least a degree of agency). Harmful bias can in part be attributable to discriminatory bias, but it can also result from the interaction of a variety of different factors

---

16 *Ibid* at 17.

17 For the definition of "harmful bias" see European Commission, "EU-U.S. Terminology and Taxonomy for Artificial Intelligence First Edition" (31 May 2023) at 11, online: <digital-strategy.ec.europa.eu> [perma.cc/M7R7-7ZPQ].

18 *Ibid* at 11 (definition of "Discrimination").

19 *Ibid* at 11 [emphasis added].

within a system which then reproduces this bias or manifests it in new ways.

The NIST AI RMF and the EU-US definitions explore a complex understanding of bias in AI and its relationship to harm and discrimination – which are not interchangeable terms. Harmful bias does not necessarily fit within the grounds of discrimination typically identified in human rights legislation. For example, a system might embed bias based on whether one is resident in a rural or urban location, creating unfair outcomes that would not be recognized as discrimination under most Canadian human rights statutes. Nevertheless, Canada's draft *Artificial Intelligence and Data Act* appears to conflate the two terms when it establishes obligations to identify and mitigate the risk of "biased output" from an AI system, defining "biased output" as:

> content that is generated, or a decision, recommendation or prediction that is made, *by an artificial intelligence system* and that adversely differentiates, directly or indirectly and without justification, in relation to an individual on one or more of the prohibited grounds of discrimination set out in section 3 of the *Canadian Human Rights Act*, or on a combination of such prohibited grounds [. . .].[20]

By pegging bias to specific categories of discrimination in human rights legislation, this approach narrows the range of bias addressed by the law. By focusing on biased *output*, it also narrows the focus of the bias inquiry and slants towards characterizations of bias as tied to machines and not the broader socio-technical context in which they are designed, deployed, and maintained.

A further concept relevant to this paper is the reasonable apprehension of bias, which is linked to fairness in administrative and judicial decision-making and addresses the decision maker's impartiality. The test is framed in terms of whether a reasonable person would consider that the decision maker was able to act fairly.[21] The apprehension of bias need not be with respect to prohibited grounds of discrimination, it could be founded on pecuniary interests or personal relationships.[22] Further, it is unnecessary to demonstrate actual bias – just a reasonable apprehension of bias; the *perception* of fairness is important to the reputation of the justice system. In the case of ADM, the concept of reasonable apprehension of bias can align somewhat with automation bias. For example, a reasonable

---

[20] *AIDA*, *supra* note 10, s 5(1) [emphasis added].

[21] *Committee for Justice and Liberty al v National Energy Board et al*, 1976 CanLII 2 (SCC) at 394.

[22] Colleen M Flood & Loren Sossin, *Administrative Law in Context*, 3rd ed (Toronto: Emond, 2018), at 282–86.

apprehension of bias might arise when a human regularly and unreflexively accepts the recommendations of a system designed to aid in decision-making.[23]

Clearly, the term 'bias' has different meanings in different contexts. There are also multiple factors that can contribute to biased outcomes in AI, and they are not limited to data and algorithms. Although *RDS* is a case about the reasonable apprehension of bias, it illustrates how the issue of reasonable apprehension of bias can become entangled in a broader discussion of discrimination. This is because specifically and pointedly, the case addresses whether a non-majoritarian understanding of the context in which a dispute arose reflects impartiality. In this sense, it resonates with the messiness of bias and discrimination in AI with which both the NIST AI RMF and the EU-US definitions struggle. The different decisions in the case are discussed in the next section.

II. *R. v. R.D.S.*

On November 10th, 1993, R.D.S., a 15-year-old youth from Nova Scotia's Black community had a run-in with a White police officer on the streets of Halifax. The officer was arresting another youth in relation to a motor vehicle theft. The court heard two versions of the event. According to the police officer, he had detained the young man and was waiting for backup when R.D.S. cut across the road and ran his bicycle up against the officer's legs. R.D.S. yelled at the officer and tried to push him away from the youth he was arresting. These facts led to R.D.S.'s arrest, who was charged with assaulting a police officer, assaulting a police officer with intent to prevent the lawful arrest of another person, and resisting arrest.[24]

According to R.D.S., he was cycling from his grandmother's house to his own when he saw a crowd forming around a police car. He rode up to the scene and recognized the detained youth. He asked him what had happened and told him that he would call the youth's mother. The police officer told R.D.S. to "shut up" or he would also be arrested. R.D.S. asked the detained youth again if he wanted him to call his mother, and the po-

---

[23] See e.g. Sancho McCann, "Discretion in the Automated Administrative State" (2023) 36:1 Can JL & Jur 171 at 187–88; see also Jennifer Raso, "AI and Administrative Law", in Florian Martin-Bariteau & Teresa Scassa, eds, *Artificial Intelligence and the Law in Canada*, 1st ed (Toronto: LexisNexis, 2021) 182 at 194.

[24] Constance Backhouse, *Reckoning with Racism, Police, Judges, and the RDS Case* (Vancouver: University of British Columbia Press, 2022), at 9; *R v RDS* CA, *supra* note 7 at 1.

lice officer put the former in a chokehold and arrested him. R.D.S. denied running into the officer with his bicycle.[25]

At trial, the officer and R.D.S. were the only witnesses. Judge Sparks found that the Crown had not met its burden of proving guilt beyond a reasonable doubt. All justices sitting in review of this case at all levels of court would have found no reasonable apprehension of bias had the decision ended there. However, Judge Sparks went on to say in oral reasons:

> The Crown says, well, why would the officer say that events occurred the way in which he has relayed them to the Court this morning. I'm not saying that the constable has misled this Court, although police officers have been known to do that in the past. And I'm not saying that the officer overreacted but certainly police officers do overreact, particularly when they're dealing with nonwhite groups. That, to me, indicates a state of mind right there that is questionable.
>
> I believe that probably the situation in this particular case is the case of a young police officer who overreacted. And I do accept the evidence of R.D.S. that he was told to shut up or he would be under arrest. That seems to be in keeping with the prevalent attitude of the day.
>
> At any rate, based upon my comments and based upon all of the evidence before the Court I have no other choice but to acquit.[26]

The Crown appealed the acquittal to the Nova Scotia Supreme Court, arguing that Judge Sparks' decision was based on considerations and findings of credibility unsupported by evidence.[27] Chief Justice Constance Glube agreed, noting that "judges must be extremely careful to avoid expressing views which do not form part of the evidence."[28] Going further, the Chief Justice applied the objective test of reasonable apprehension of bias, "whether a reasonable right-minded person with knowledge of all the facts would conclude that the judge's impartiality might reasonably be questioned."[29] She concluded that "in spite of the thorough review of the facts and the finding on credibility, the two paragraphs at the end of the decision lead to the conclusion that a reasonable apprehension of bias exists."[30]

---

25 Backhouse, *supra* note 24 at 17; *R v RDS* CA, *supra* note 7 at 2.

26 *R v RDS* CA, *supra* note 7 at 3.

27 *R v RDS* SC, supra note 7 at para 6.

28 *Ibid* at para 25.

29 *Ibid* at para 26.

30 *Ibid*.

The majority of a three-judge panel of the Nova Scotia Court of Appeal confirmed this decision. They found it "apparent" that Judge Sparks based her decision "at least in part, on her general comments with respect to the police."[31] In response to arguments by counsel for R.D.S. that the comments "merely reflect an unfortunate social reality,"[32] the majority noted that the real issue was whether "the Youth Court Judge, considered matters *not in evidence* in arriving at her critical findings of credibility, and hence, acquittal."[33] The majority criticized her "unfortunate use of these generalizations," and found a reasonable apprehension of bias.[34]

Justice Freeman, in his dissent, noted that assessment of credibility is "a notoriously difficult and inexact exercise in adjudication in which the judge's whole background experience plays a role in the assessment of demeanour and other intangibles."[35] He highlighted the "racially charged" nature of the case, stating that "Judge Sparks was under a duty to be sensitive to the nuances and implications, and to rely on her own common sense which is necessarily informed by her own experience and understanding."[36] Rather than finding that Judge Sparks' comments were addressed to evidence not before the court, he treated the matter as one of determining credibility, which "draws upon all of the judge's wisdom and experience."[37] He observed that while Judge Sparks' comments could have been clearer, they did not raise a reasonable apprehension of bias.

On further appeal, a majority of the Supreme Court of Canada ruled that Judge Sparks' comments did not raise a reasonable apprehension of bias. Two of the six majority judges did so with reservations. Justice Cory, writing for himself and for Justice Iacobucci found that a judge is "obviously permitted to use common sense and wisdom gained from personal experience in observing and judging the trustworthiness of a particular witness on the basis of factors such as testimony and demeanour." However, a judge "must avoid judging the credibility of the witness on the basis of generalizations or upon matters that were not in evidence."[38] Alt-

---

31 *R v RDS* CA, *supra* note 7 at 10.

32 *Ibid*.

33 *Ibid* [emphasis added].

34 *Ibid* at 11.

35 *Ibid* at 15.

36 *Ibid*.

37 *Ibid* at 17.

38 *R v RDS* SCC, *supra* note 7 at para 129.

hough he found no reasonable apprehension of bias, Justice Cory characterized Judge Sparks' remarks as "worrisome"[39] and "troubling."[40]

Justices L'Heureux-Dubé and McLachlin wrote a separate opinion, with which Justices LaForest and Gonthier concurred. They agreed that there was no reasonable apprehension of bias, but disagreed with the conclusion that the remarks were inappropriate. They maintained the importance of judges bringing their experience to their role, and emphasized that while impartiality was important, judges were not to act as "neutral ciphers."[41] They also stressed the importance of context to judicial decision-making. They noted that systemic racism was a reality in Nova Scotia, stating: "[t]he reasonable person is cognizant of the racial dynamics in the local community, and, as a member of the Canadian community, is supportive of the principles of equality."[42] Further, they observed that an awareness of context is not a lack of neutrality; rather, it is "consistent with the highest tradition of judicial impartiality."[43] On reviewing Judge Sparks' comments, they found nothing to indicate pre-judgment. Rather, they ascertained that Judge Sparks' comments showed she had "approached the case with an open mind, used her experience and knowledge of the community to achieve an understanding of the reality of the case, and applied the fundamental principle of proof beyond a reasonable doubt."[44]

Three dissenting justices, under Justice Major's pen, found that the facts raised a reasonable apprehension of bias. Justice Major stated: "A fair trial is one that is based on the law, the outcome of which is determined by evidence, free of bias, real or apprehended."[45] He concluded that the decision was not based on the evidence before the court, but on "something else."[46] He went so far as to state that Judge Sparks had stereotyped all police officers as liars and racists, declaring: "It would be stereotypical reasoning to conclude that, since society is racist, and in effect, tells minorities to 'shut up,' we should infer that *this* police officer told *this* appel-

---

39 *Ibid* at para 152.

40 *Ibid* at para 151. Note that Dianne Pothier is critical of Justice Cory's approach on the basis that it emphasizes formal rather than substantive equality (see Richard F Devlin & Dianne Pothier, "Redressing the Imbalances: Rethinking the Judicial Role After *R. v. R.D.S.*" (1999-2000) 31:1 Ottawa L Rev 1 at 31).

41 *R v RDS* SCC, *supra* note 7 at para 38.

42 *Ibid* at para 48.

43 *Ibid* at para 49.

44 *Ibid* at para 59. For a discussion and critique of the different reasons in this case, see Devlin & Pothier, *supra* note 40.

45 *R v RDS* SCC, *supra* note 7 at para 3.

46 *Ibid*.

lant minority youth to 'shut up'."[47] It was an error in law for Judge Sparks "to infer that based on her general view of the police or society,"[48] the police officer's actions and testimony were informed by racism. According to Justice Major, "[l]ife experience is not a substitute for evidence."[49]

Although *RDS* was framed as an issue of reasonable apprehension of bias or judicial impartiality, the impartiality issue really turned on whether a racialized judge could publicly acknowledge the perspective that informed her assessment of the law and facts. Since her perspective was non-majoritarian, many judges involved did not consider it as "neutral" or impartial. This inability to distinguish between different lived experiences and bias—or the tendency to characterize non-mainstream perceptions as biased —raises important questions about how bias will be recognized and identified in the AI context, and by whom.

III. Themes from *R. v. R.D.S.*

This section explores four ways in which *RDS* is important to our understanding of discriminatory bias in automated decision making. It considers the tension between fact and opinion, issues of transparency and explainability, biased input and output, and the human-in-the-loop.

A. Fact is a matter of opinion?

A key issue in *RDS* is how to distinguish between "objective" fact and evidence on one hand, and "subjective" stereotype or opinion on the other. One of the asserted virtues of AI is its elimination of subjective opinion and focuses on objective data for decision-making.[50] Yet, although there is a tendency in some quarters to treat data as a kind of absolute truth,[51] critical data scholars remind us that data are not neutral. For example, Rob Kitchin describes data as "capta," meaning "those units of data that

---

47 *Ibid* at para 9 [emphasis in original].

48 *Ibid* at para 10.

49 *Ibid* at para 13.

50 See e.g. Eric Colson, "What AI-Driven Decision Making Looks Like", *Harvard Business Review*, (8 July 2019), online: <hbr.org> [perma.cc/KZ9H-GGQA]; Bruno Lepri et al, "Fair, Transparent, and Accountable Algorithmic Decision-Making Processes" (2018) 31:4 Philosophy & Tech 611 at 622.

51 For example, in *Ewert v Canada*, 2018 SCC 30 at para 41, the majority observes "the Crown took the position that actuarial tests are an important tool *because the information derived from them is objective* and thus mitigates against bias in subjective clinical assessments" [emphasis added].

have been selected and harvested from the sum of all potential data."[52] Implicit are the choices that went into the decision to capture particular data about a specific subject matter.

When it comes to data about some groups or communities, there may be further issues: an absence of some key data and a lack of involvement of the community in how data are collected or used. The lack of adequate data about racialized communities, for example, is such a problem that Ontario and British Columbia have passed laws seeking to capture more data about these communities while ensuring that they will not be used in harmful ways.[53] Invisibility in data leads to poor outcomes in a data-driven society; yet visibility without control or input is dangerous.[54]

An important issue in *RDS* is what counts as fact in the first place. There are clear differences between the appellate justices as to whether Judge Sparks made findings of fact unsupported by evidence, or findings of credibility based on life experience that led to conclusions about what was or was not proven as fact (i.e., what happened). For the judges who found a reasonable apprehension of bias, Judge Sparks could only have made racism an issue if evidence about it were adduced and linked to a legal argument. For other judges, life experience supports assessments of credibility which can lead to conclusions about the facts. Thus, racism is either part of the life experience of a judge that informs her determinations of fact, or it is itself a social fact that must be proven before it can be relevant to a decision. By contrast, the role of dominant perspectives in shaping what constitute legal facts typically goes unquestioned.

---

[52] Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences* (London, UK: SAGE Publications Ltd, 2014) at 2.

[53] See *Anti-Racism Act*, 2017, SO 2017, c 15; *Anti-Racism Data Act*, SBC 2022, c 18. In Nova Scotia, the *Dismantling Racism and Hate Act*, SNS 2022, c 3 attempts to address systemic racism in the province. It addresses data equity by requiring the Minister, in s 11(1), to establish data standards for the collection of data that can be used to "identify, monitor and address systemic hate, inequity and racism."

[54] The Indigenous data sovereignty movement emphasizes the importance of both control over data and a defining role for Indigenous peoples in determining the purposes and boundaries of data collection. See e.g. First Nations Information Governance Centre, "Ownership, Control, Access and Possession (OCAP™): The Path to First Nations Information Governance", (23 May 2014) at 11–13, online (pdf): <fnigc.ca> [perma.cc/VXP2-XA8C]; Maggie Walter & Stephanie Russo Carroll, "Indigenous Data Sovereignty, Governance and the Link to Indigenous Policy" in Maggie Walter et al, eds, *Indigenous Data Sovereignty and Policy* (Abingdon: Routledge, 2021) 1 at 2–3. Similar concepts of control are central in the call for more community-based control and input regarding the health data of Black communities in Ontario (see Black Health Equity Working Group, "Engagement, Governance, Access, and Protection (EGAP): A Data Governance Framework for Health Data Collected from Black Communities in Ontario", (2021) at 11, online (pdf): <blackhealthequity.ca> [perma.cc/L2KZ-HUZ]).

This dispute over how facts are made is instructive in the AI context as it makes explicit how human judgment shapes the data on which we rely. It also challenges the neutrality of facts and data, and centralizes the issue of who gets to determine what constitute facts. The NIST Framework identifies human cognitive bias as an important factor in AI bias, and the EU-US definitions also make it clear that bias can be found not just in data assembled and curated by humans, but in human decisions and processes across the AI lifecycle. The issues that surfaced in *RDS* around what is fact and what is opinion are also implicit in the building and operation of AI systems. Rather than purely technical tools, AI systems are complex socio-technical systems that are deeply embedded within a framework that includes people, processes, rules, and norms. The challenge is how to surface, query, and challenge these issues in ADS.

B. Transparency and Explainability

Transparency and explainability are identified as core values in ethical AI.[55] In the US NIST AI RMF, explainability is defined as "a representation of the mechanisms underlying AI systems' operation".[56] In other words, interpretability refers to a kind of cause-and-effect logic. Transparency aids in scrutiny, but as the NIST AI RMF points out, "[a] transparent system is not necessarily an accurate, privacy-enhanced, secure, or fair system."[57]

Rights to an explanation of ADM are often quite limited. For example, in Canada's proposed *Consumer Privacy Protection Act* (CPPA), the right to an explanation of a "prediction, recommendation or decision" is available only where the decision may have a "significant impact" on the data-subject.[58] The content of an explanation includes "the type of personal information that was used to make the prediction, recommendation or decision, the source of the information and the reasons or principal factors

---

[55] See e.g. OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (2024) at 4, online (pdf): <oecd.ai> [perma.cc/B4RW-Y3FW]; UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (23 November 2021) at 22, online (pdf): <unesdoc.unesco.org> [perma.cc/X3FH-7GNC]; NIST, *AI RMF*, *supra* note 10 at 15–16.

[56] NIST, *AI RMF*, *supra* note 10 at 16 distinguishes between explainability and interpretability, with the latter referring to "the meaning of AI systems' output in the context of its designed functional purpose."

[57] *Ibid* at 16.

[58] Bill C-27, *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*, 1st Session, 44th Parliament, cl 63(3) (first reading 16 June 2022).

that led to the prediction, recommendation or decision."[59] Under article 13(2)(f) of the *General Data Protection Regulation* (GDPR), data subjects have the right to "meaningful information about the logic involved as well as the significance and the envisaged consequences of such processing for the data subject".[60] Explainability operates on a systemic rather than an individual basis, in large part because it is presumed that the system processes data in an objective way—outcomes are adequately explained by a description of parameters and inputs.[61]

The right to reasons in administrative law is a principle of administrative fairness, yet the extent of this right varies considerably. For routine administrative decisions of little consequence, general and formulaic explanations will suffice. More detailed reasons may only be required where assessments of credibility are made, the decision has an important impact on the individual, or there is a statutory right of appeal.[62] The form and extent of these reasons may also vary.[63]

In the AI context, basic rights to an explanation or to interpretability may not help in exposing bias or discrimination. In *RDS*, a majority of justices across all levels of appeal found that it would have been better for Judge Sparks to say less rather than more in reaching her decision. For example, Justice Glube stated that "judges must be extremely careful to avoid expressing views which do not form part of the evidence,"[64] suggesting that it is the expressing, and not the holding of the views that matters. Justice Cory, for the majority of the Supreme Court of Canada agreed that "if the decision had ended after the general review of the evidence and the resulting assessments of credibility,"[65] there would have

---

59 *Ibid* at cl 63(4).

60 EU, *Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)* [2017] OJ, L 119/1 at art 13(2)(f) [GDPR]; see also *ibid* at cls 14(2)(g), 15(1)(h).

61 See Lilian Edwards & Michael Veale, "Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'?" (2018) 16:3 IEEE Security & Privacy (2018) 44. Edwards and Veale note that as framed, "it is uncertain if the right is only to a general explanation of the model of the system as a whole (model-based explanation), rather than an explanation of how a decision was made based on that particular data subject's particular facts (subject-based explanation)" (48).

62 *Baker v Canada (Minister of Citizenship and Immigration)*, 1999 CanLII 699 (SCC) at para 43; *Suresh v Canada (Minister of Citizenship and Immigration),* 2002 SCC 1 at para 126. See also Raso, *supra* note 23 at 192.

63 *Baker*, *supra* note 62 at para 43; Raso, *supra* note 23 at 192.

64 *R v RDS* SC, *supra* note 7 at para 25.

65 *R v RDS* SCC, *supra* note 7 at para 145.

been no basis on which to impugn it. Judge Sparks' principal problem, it would seem, was that she deviated from the standard script and provided insight into her thought process. Formalized rights to an explanation in ADM may ultimately offer little help in unpacking—or challenging—the assumptions and human-cognitive choices in the system design or data classification that led to outcomes.

C. Biased Input and Biased Output

Although *RDS* dates to the 1990's, we do not have to look very far into the past to find examples of how the existence of bias and discrimination are still contested in our society. Acceptance of systemic bias is particularly challenging. For example, in 2021, the Premier of Quebec argued that there was no systemic bias in Quebec, relying on his own interpretation of a dictionary definition of 'systemic.'[66] In 2020, the Commissioner of the Royal Canadian Mounted Police also denied the presence of systemic racism in that force.[67] Taking a data-driven approach, a recent think tank report used statistical methods to contest the existence of systemic discrimination in Canada.[68] The fundamental nature of some human rights claims are also flat-out contested. For example, equality rights claims from the LGBTQ+ communities have been consistently resisted by those claiming that such rights conflict with their religious views.[69]

These examples suggest that addressing bias and discrimination in AI may be more complex than typically presented, even if it is identified as an ethical and legal imperative.[70] Different approaches are being devel-

---

66 René Bruemmer, "After Echaquan report, Legault repeats there is no systemic racism in Quebec", *Montreal Gazette* (5 October 2021), online: <montrealgazette.com> [perma.cc/JKA4-QB5N].

67 Daniel Leblanc & Kristy Kirkup, "RCMP commissioner 'struggles' with definition of systemic racism, but denies its presence in the organization", *Globe and Mail* (last updated 11 June 2020), online: <theglobeandmail.com> [perma.cc/DUH3-WGEE]. Note that Commissioner Lucki did subsequently acknowledge the existence of systemic discrimination in the RCMP (see John Paul Tasker, "Systemic racism exists in the RCMP, Commissioner Brenda Lucki says", *CBC News* (last updated 13 June 2020), online: <cbc.ca> [perma.cc/PKH5-5VL6]).

68 Matthew Lau, "Systemic racism claims in Canada: A fact-based analysis" (30 October 2023), online: <aristotlefoundation.org> [perma.cc/VM6G-37N4].

69 See e.g. Human Rights Watch, "United States: State Laws Threaten LGBT Equality" (19 February 2018), online: <hrw.org> [perma.cc/E688-LLEF]. Note that in *R v RDS*, the dissenting justices at the Supreme Court of Canada 'flip' the narrative, characterizing Judge Sparks' comments as stereotyping police as liars and racists.

70 European Commission, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts" (4 April 2021), online: <eur-lex.europa.eu> [perma.cc/VD2Z-2ETV] [EU AI Act]; AIDA, *supra* note 10.

oped to identify and monitor for bias and discrimination in AI. These means include techniques to improve data quality, and to disrupt factors that may lead to biased correlations.[71] Canada's AIDA requires the person responsible for a high-impact AI system to "establish measures to identify, assess and mitigate the risks of harm or biased output that could result from the use of the system."[72] One mitigation measure could be determining if data are representative of the relevant community to which the decision system will apply and making necessary corrections if the data are biased. Yet, as noted earlier, data curation alone will not suffice. Problems may arise from how training data were classified or weighed.[73] There may also be issues around the decision to use ADM in this context rather than human decision-makers. Furthermore, issues may arise around how any human-in-the-loop interacts with and responds to the AI system.

Countering the concerns that AI systems can perpetuate bias, some argue that ADM has potential to greatly improve notoriously flawed human decision-making.[74] Some human decision-makers might not only be biased, but they might also have techniques that allow them to mask it (for example, by emphasizing other factors in reasons for decision). Risk mitigation measures that include scrutiny of training data and ongoing monitoring of decision-making outputs in ADS have the potential to identify bias and to correct it, thus, in theory, making decision-making fairer and more impartial.[75] These are serious arguments. Yet, they depend to some extent on the idea that the answers lie in better data and algorithms. The more complex definitions of harm and bias discussed earlier make it clear that humans and their decision-making processes are still deeply embedded in the choice, design, and implementation of ADS. If approaches to bias and discrimination in AI are reduced to an assessment and modification of algorithms and data, these machine-focused bias solu-

---

71 Zhisheng Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices" (2023) 10:567 Humanities & Soc Sciences Communications 1; see also Lobel, *supra* note 8, ch 2.

72 AIDA, *supra* note 10, s 8.

73 See e.g. Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven: Yale University Press, 2021) at 135.

74 See e.g. Lobel, *supra* note 8 at 5–12, 77–82.

75 See e.g. the prescribed approaches in the Treasury Board of Canada, *Directive on Automated Decision-Making* (Directive) (Ottawa: Treasury Board of Canada Secretariat, 2021) last updated on 25 April 2023, online:<tbs-sct.canada.ca> [perma.cc/727A-EQ3P] [DADM]; or the requirements in the AIDA *supra* note 10, ss 8–9.

tions may short-circuit the complex and difficult discussions needed to address broader manifestations of bias in AI.[76]

D. The human-in-the-loop

In *RDS* the different justices assess the same paragraphs of Judge Spark's decision looking to see if there is a reasonable apprehension of bias. Their conclusions are as much about what each perceives as meeting the legal test for bias as they are about their understanding of how lived experience shapes the interpretation of evidence. In *RDS*, we see how the dominant group's lived experience is the largely unquestioned norm. In this way, the case centres the role of the human decision-maker and the relevance of identity in the decision-making process.

Where concerns are expressed about ADM, a "human-in-the-loop" is typically proposed to ensure a degree of actual or potential human participation in the decision-making process. The EU's GDPR implicitly requires a human-in-the-loop for ADM, providing that "[t]he data subject shall have the right not to be subject to a decision based solely on automated processing [...]."[77] Article 14 of the EU AI Act, requires human oversight for high-risk systems, stating:

> Human oversight shall aim at preventing or minimising the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of other requirements set out in this Chapter.[78]

The human-in-the-loop humanizes the process and provides a backstop against harmful and discriminatory bias that may have escaped technological risk mitigation measures. Notably, Canada's AIDA and proposed CPPA do not require a human-in-the-loop for automated decision systems,[79] although the federal *Directive on Automated Decision-Making* pro-

---

[76] Nicol Turner Lee et al recommend a multi-pronged approach to eliminating bias that includes "the development of a bias impact statement, inclusive design principles, and cross-functional work teams." They also recommend updating anti-discrimination laws to apply to digital contexts. Their proposed approach is aptly complex and multi-faceted (Nicol Turner Lee, Paul Resnick & Genie Barton, "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms" (22 May 2019), online: <brookings.edu> [perma.cc/S74U-LGE5]).

[77] GDPR, *supra* note 60 at art 22(1). Note that this right is subject to exceptions.

[78] EU AI Act, *supra* note 70 at art 14.

[79] In comments on a similar provision in the predecessor to Bill C-27, former Privacy Commissioner Therrien recommended that the CPPA include a right to context an automated decision (see Office of the Privacy Commissioner of Canada, *Submission of the*

vides that in cases of high-impact ADM, a final decision must be made by a human.[80]

In spite of the backstop role of the human-in-the-loop, there is little consideration of who the human-in-the-loop is or their status in relation to the decision-making process.[81] The human-in-the-loop may not be a decision-maker in the more formal sense of a tribunal member or a judge. There are no examples of provisions for an appointment process that might speak to independence or impartiality, nor is it clear what, if any, consideration will be given to the person's background or experience. It is similarly unclear whether a human-in-the-loop will be someone with understanding of the technical features of the system, or a person trained in the policy behind the decision-making system, or even in ethics. The genericness of humans-in-the-loop is particularly interesting considering *RDS*, where the identification and experience of the decision-maker(s) at every stage was a central factor.[82]

Certainly, the extent to which judges are representative of Canadian society remains an issue that is important to equity and fairness in the justice system.[83] Moreover, just as there is underrepresentation in adjudicative roles, there is substantial underrepresentation of many groups among those involved in the design and development of AI—including women and racialized persons.[84] In this context, it is uncertain whether a person affected by ADM will have any means of knowing the identity of the human-in-the-loop who reviewed or participated in the decision. Neither is it evident how humans-in-the-loop will be assessed for automation bias. It is possible that in some cases, their own performance will be monitored by an automated system, which could impact impartiality if re-

---

Office of the Privacy Commissioner of Canada on Bill C-11, the Digital Charter Implementation Act, 2020 (Ottawa: Office of the Privacy Commissioner of Canada, 2021) at recommendation 28, online: <priv.gc.ca> [perma.cc/AZN3-KSFA]).

80 DADM, *supra* note 75.

81 For an examination of some of the problems with the human-in-the-loop concept (see Rebecca Crootof et al, "Humans in the Loop" (2023) 76:2 Vand L Rev 429 at 436–37).

82 In her history of *R v RDS* SCC, Constance Backhouse observes that intersectionality played a role, suggesting that Judge Sparks' gender, combined with her race, impacted how her words were interpreted (see Backhouse, *supra* note 24 at 118–23).

83 See e.g. Erin Crandall, "A Reflection of Canadian Society? An Analysis of Federal Appointments to Provincial Superior Courts by the Liberal Government of Justin Trudeau" (2022) 45:2 Dal LJ 359; *Equality in Judicial Appointments*, Canadian Bar Association (2013) Res 13-04-A, online: <cba.org> [perma.cc/7KJF-H6C4].

84 See e.g. Council of Canadian Academies, *Leaps and Boundaries: The Expert Panel on Artificial Intelligence for Science and Engineering* (Ottawa: Council of Canadian Academies, 2022) at 67–68, online: <cca-reports.ca> [perma.cc/V6VH-H6PK].

peated divergence from AI recommendations is seen as anomalous behavior.

Perhaps most importantly, though, it remains uncertain whether the role of the human-in-the-loop is just to be a so-called 'neutral' check on the operations of ADS or whether this is also a context in which we continue to seek diversity in perspectives to shape and inform decision-making. The authors of a paper on intersectionality and AI bias call for a different and more inclusive approach to assessing fairness in AI, including "a widening of AI fairness practice by centering marginalized people and valorizing critical knowledge production that makes room for their voices."[85] This is an encouraging, although rare articulation of this view. The lack of attention to the identity of the human-in-the-loop – in other words, their presumed neutrality – deeply resonates with the issues of identity and bias that are surfaced by *RDS*.

Conclusion

Risk regulation, the dominant paradigm for AI governance, is premised on the existence of risks that must be mitigated. Such risks include harmful bias and discrimination, which will be disproportionately borne by those who have experienced generations of discrimination, compounding existing inequality. Furthermore, although bias and discrimination are often presented as issues of data quality or flawed assumptions in algorithms, *RDS* teaches us that the problems are more complex than merely biased or incomplete data. There may be fundamental differences as to how we are prepared to understand or interpret the data, how we build the systems to process the data, how we adopt, implement and oversee systems, and who is engaged in these processes. While the NIST AI RMF and the EU-US AI definitions attempt to capture this broader understanding of how bias and discrimination may be manifested in ADM, this approach is less evident in Canada. In all contexts, there is a real risk that risk-mitigation measures will be reduced to automated assessments of outputs and enhanced data curation. Even though these are important activities, they are not sufficient. Just as the problems are not solely in the machines, neither are the solutions.

More fulsome approaches to bias and discrimination in AI are not limited to issues of data quality or coded assumptions; they go so far as to include the very choices that are made about how to deploy AI and in what

---

[85] Anaelia Ovalle et al, "Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness" (Paper delivered at the Sixth AAAI/ACM Conference on AI, Ethics, and Society, Montréal, 8–10 August 2023) 496, online: <dl.acm.org> [perma.cc/ZNN5-55J7].

contexts. *RDS* reminds us that very experienced, highly-trained and well-paid and respected members of society can have profoundly different opinions about the constitution of facts and the existence of bias. It is a reminder that bias and discrimination in AI systems are fundamentally human issues, and artificial intelligence is still a fundamentally human technology from its inception to its deployment. This reasoning suggests that we have much work to do—and much more challenging and complex work at that—in order to address bias and discrimination in AI.