International Journal of Refugee Law, 2022, Vol XX, No XX, 1–32
https://doi.org/10.1093/ijrl/eeac040

OXFORD

What Can Artificial Intelligence Do for Refugee Status Determination? A Proposal for Removing Subjective Fear

Niamh Kinchin*[*] and Davoud Mougouei[†]

ABSTRACT

The drive for innovation, efficiency, and cost-effectiveness has seen governments increasingly turn to artificial intelligence (AI) to enhance their operations. The significant growth in the use of AI mechanisms in the areas of migration and border control makes the potential for its application to the process of refugee status determination (RSD), which is burdened by delay and heavy caseloads, a very real possibility. AI may have a role to play in supporting decision makers to assess the credibility of asylum seekers, as long as it is understood as a component of the humanitarian context. This article argues that AI will only benefit refugees if it does not replicate the problems of the current system. Credibility assessments, a central element of RSD, are flawed because the bipartite standard of a 'well-founded fear of being persecuted' involves consideration of a claimant's subjective fearfulness and the objective validation of that fear. Subjective fear imposes an additional burden on the refugee, and the 'objective' language of credibility indicators does not prevent the challenges decision makers face in assessing the credibility of other humans when external, but largely unseen, factors such as memory, trauma, and bias, are present.

Viewing the use of AI in RSD as part of the digital transformation of the refugee regime forces us to consider how it may affect decision-making efficiencies, as well as its impact(s) on refugees. Assessments of harm and benefit cannot be disentangled from the challenges AI is being tasked to address. Through an analysis of algorithmic decision making, predictive analysis, biometrics, automated credibility assessments, and digital forensics, this article reveals the risks and opportunities involved in the application of AI in RSD. On the one hand, AI's potential to produce greater standardization, to mine and parse large amounts of data, and to address bias, holds significant possibility for increased consistency, improved fact-finding, and corroboration. On the other hand, machines may end up replicating and manifesting the unconscious biases and assumptions of their

* Senior Lecturer, School of Law, University of Wollongong, NSW, Australia. Email: nkinchin@uow.edu.au.

† Lecturer, School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, QLD, Australia.

human developers, and AI has a limited ability to read emotions and process impacts on memory. The prospective nature of a well-founded fear is counter-intuitive if algorithms learn based on training data that is historical, and an increased ability to corroborate facts may shift the burden of proof to the asylum seeker. Breaches of data protection regulations and human rights loom large. The potential application of AI to RSD reveals flaws in refugee credibility assessments that stem from the need to assess subjective fear. If the use of AI in RSD is to become an effective and ethical form of humanitarian tech, the 'well-founded fear of being persecuted' standard should be based on objective risk only.

1. INTRODUCTION

The perennial search for greater accuracy, consistency, cost-effectiveness, and efficiency in administrative decision making[1] has prompted governments to increasingly turn to artificial intelligence (AI).[2] Credibility assessments, which are the most multifaceted, inefficient, and uncertain element of refugee status determination (RSD), offer an ideal canvas for AI development. Wide recognition of deficiencies in credibility assessments,[3] largely attributable to the need for refugees to demonstrate that they are subjectively fearful, raises questions about the feasibility and fairness of credibility assessments in their current form. Could the integration of AI into RSD, as a complementary or alternative method to human credibility assessments, create an opportunity for greater efficiency and accuracy? Or would the use of AI mechanisms pose new risks to the fairness of a process defined by human vulnerability? In seeking answers to these questions, the flaws of credibility assessments are laid bare, revealing that the potential beneficence of AI in RSD hinges on the removal of the subjective element of the 'well-founded fear' standard.

RSD is the decision-making process that States and the United Nations High Commissioner for Refugees (UNHCR) use to determine whether an asylum seeker meets the criteria to be recognized as a refugee according to the definition in the Convention relating to the Status of Refugees (Refugee Convention).[4] RSD is a

[1] Commonwealth Ombudsman, 'Automated Assistance in Administrative Decision-Making Better Practice Guide' (2007) <https://www.ombudsman.gov.au/publications/better-practice-guides/automated-decision-guide> accessed 25 August 2022.

[2] Artificial intelligence (AI) describes the theory and development of 'computational agents that act intelligently', or computer systems that perform tasks usually performed by humans, which require some human or cognitive thought processes: see David L Poole and Alan K Mackworth, *Artificial Intelligence: Foundations of Computational Agents* (2nd edn, Cambridge University Press 2017) 3.

[3] See eg Trish Luker, 'Decision-Making Conditioned by Radical Uncertainty: Credibility Assessment at the Australian Refugee Review Tribunal' (2012) 25 International Journal of Refugee Law 502; Hilary Evans Cameron, 'Refugee Status Determination and the Limits of Memory' (2010) 22 International Journal of Refugee Law 469; Laura Smith-Khan, 'Why Refugee Visa Credibility Assessments Lack Credibility: A Critical Discourse Analysis' (2019) 24 Griffith Law Review 406.

[4] Convention relating to the Status of Refugees (adopted 28 July 1951, entered into force 22 April 1954) 189 UNTS 137 (Refugee Convention) art 1A(2), read in conjunction with the Protocol relating to the Status of Refugees (adopted 31 January 1967, entered into force 4 October 1967) 606 UNTS 267 (Protocol).

complex, expensive process, burdened by delay and prone to uncertainty. Whilst waiting times for decisions differ, most country systems and UNHCR are experiencing increasing delays.[5] Although the potential for integrating AI into RSD has received only cursory attention to date, the significant and continuing growth of AI in government programmes – including migration[6] and border control – as well as its expanding use within the United Nations (UN) system,[7] makes the potential for its application to RSD a real possibility for authorities seeking to better manage heavy caseloads.

Part 2 contextualizes AI in RSD as a form of 'humanitarian tech'. Humanitarian tech, a burgeoning aspect of humanitarian design, comprises 'the technologies used to collect, process, and analyse information that may contribute to improve the livelihoods of people affected by a devastating event'.[8] This part argues that while humanitarian

[5] For example, in August 2019, asylum seekers in the United Kingdom (UK) were reportedly waiting more than six months for decisions, a 58% increase on the previous year: see May Bulman, 'Asylum Waiting Times at Record High as Thousands "Left in Limbo"' *The Independent* (22 August 2019) <https://www.independent.co.uk/news/uk/home-news/asylum-seekers-waiting-times-home-office-immigration-a9075256.html> accessed 26 March 2021. UNHCR has also reported sizeable backlogs in its own RSD practice: see Brian Barbour, 'Refugee Status Determination Backlog Prevention and Reduction', UNHCR Legal and Protection Policy Research Series, PPLA/2018/03.

[6] The Canadian government has been developing a system of 'predictive analytics' to automate certain activities conducted by immigration officials and to support the evaluation of some immigrant and visitor applications. It has also instituted a pilot program titled 'Artificial Intelligence Solution' for which it sought input from industry vendors about how machine learning-powered solutions that leverage data-driven insights might assist and inform: legal research and the development of legal advice and legal risk assessments; the prediction of outcomes in litigation; and trend analysis in litigation. As part of the same program, the government has sought to 'explore the possibility of whether or not the machine learning powered solution(s) could also be used by front-end [immigration, refugee, and citizenship] administrative decision-makers across the domestic and international networks to aid in their assessment of the merits of an application before decisions are finalized': see Public Works and Government Services Canada, 'Artificial Intelligence Solution (B8607-180311/A)' (Tender Notice, 2018) <https://buyandsell.gc.ca/procurement-data/tender-notice/PW-EE-017-33462> accessed 2 January 2020. In the UK, a media report in June 2019 revealed that the Home Office is using an AI system that streams visa applications using a colour code system (ie red, green, yellow) according to the category of risk: see Helen Warrell, 'Home Office under Fire for Using Secretive Visa Algorithm' *Financial Times* (10 June 2019) <https://www.ft.com/content/0206dd56-87b0-11e9-a028-86cea8523dc2> accessed 26 March 2021. The then Immigration Minister was quoted as saying that '[t]he new streamlined service will make the visa application process quicker and easier to access than ever before for people in the UK, through increasing the use of digital services': see 'Sopra Steria Has Been Awarded a New UKVI Contract' (*Sopra Steria*, 17 May 2018) <https://www.soprasteria.com/industries/government/sopra-steria-has-been-awarded-a-new-ukvi-contract> accessed 26 March 2021.

[7] International Telecommunication Union, 'United Nations Activities on Artificial Intelligence (AI)' (2021) <https://www.itu.int/pub/S-GEN-UNACT-2021> accessed 12 September 2022.

[8] Sonia Camacho, Andrea Herrera, and Andreas Barrios, 'Refugees and Social Inclusion: The Role of Humanitarian Information Technologies' in Sebastián Villa and others (eds), *Decision-Making in Humanitarian Operations* (Palgrave Macmillan 2019) 104.

tech can be presented as a value-neutral means to a beneficial end,[9] it can also be a harbinger of harm. An analysis of the opportunities and risks of using AI in RSD must therefore be viewed through a socio-legal perspective that balances function with dysfunction and recognizes that new technologies can create risks for people with limited agency. Part 3 examines credibility assessments in their current form by revealing the 'human problem' of credibility indicators and, in doing so, argues the case against subjective fear. Within the 'human problem' of credibility assessments is an opportunity to consider how AI might create a means to reconceptualize 'well-founded fear' as an objective standard only. Part 4 explores the context, form, and function of AI to identify the risks and opportunities it may bring to the refugee context. Part 5 briefly summarizes why the future application of AI in RSD as a legitimate techno-human process depends upon the removal of the subjective element of the 'well-founded fear' standard.

2. THE USE OF ARTIFICIAL INTELLIGENCE IN REFUGEE STATUS DETERMINATION AS HUMANITARIAN TECH: FUNCTION AND DYSFUNCTION

At the core of 'humanitarian design' – defined as 'products, services, or systems for populations affected by natural and/or human-made disasters'[10] – is a sense that it can 'make a difference' and 'turn a crisis into an opportunity'.[11] Humanitarian tech is often presented as exclusively beneficial for its 'subjects' in that it is designed to address or remedy 'humanitarian issues', such as social inclusion,[12] rehabilitation,[13] refugee health outcomes,[14] aid-related information analysis,[15] and permanent settlement.[16] Technology can be a form of empowerment for vulnerable populations when it lowers

[9] See eg Sheraz Memon, 'How Technology Can Expedite Essential Humanitarian Aid' (*Forbes*, 26 August 2022) <https://www.forbes.com/sites/servicenow/2022/08/26/how-technology-can-expedite-essential-humanitarian-aid/?sh=7f9d0176a380> accessed 12 September 2022.

[10] Brita Fladvad Nielsen, 'Humanitarian Design' in Antonio De Lauri (ed), *Humanitarianism: Keywords* (Brill 2020) 97.

[11] Mahmoud Keshavarz, 'Violent Compassions: Humanitarian Design and the Politics of Borders' (2020) 36 Design Issues 20, 20.

[12] Helen C Leligou and others, 'Embracing Novel ICT Technologies to Support the Journey from Camp to Job' (IEEE International Conference on Multimedia and Expo Workshops, London, 6–10 July 2020).

[13] Joydhriti Choudhury and others, 'Predicting Priority Needs for Rehabilitation of Refugees Based on Machine Learning Techniques from Monitoring Data of Rohingya Refugees in Bangladesh' (IEEE Region 10 Symposium, Dhaka, 5–7 June 2020).

[14] Nick Drydakis, 'Mobile Applications Aiming to Facilitate Immigrants' Societal Integration and Overall Level of Integration, Health and Mental Health: Does Artificial Intelligence Enhance Outcomes?' (2021) 117 Computers in Human Behavior <https://www.sciencedirect.com/science/article/pii/S0747563220304088> accessed 25 August 2022.

[15] Małgorzata Jenerowicz and others, 'Multifractality in Humanitarian Applications: A Case Study of Internally Displaced Persons/Refugee Camps' (2019) 12 IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing 4438.

[16] Kirk Bansak and others, 'Improving Refugee Integration through Data-Driven Algorithmic Assignment' (2018) 359 Science 325.

the barriers to participation in political processes and local decision making, increases transparency, and enables horizontal flows of communication. Some humanitarian tech actively seeks to facilitate design engagement with its subjects[17] and vulnerable populations can share information and communicate directly instead of being dependent on intermediaries, who have 'their own biases and insecurity regarding the sharing of power'.[18]

However, the diffusion of any innovation will result in both desirable or functional consequences and unforeseen or dysfunctional effects.[19] Technology is not neutral, nor is it a passive means through which predefined ends are carried out.[20] Drawing on Grégoire Chamayou's conception of 'vile bodies' (that is, bodies that society accords lesser value, such as slaves or prisoners), Jacobsen and Fast argue that humanitarian tech experiments with subjects in a way that blurs care and control.[21] Refugees can be singled out by States 'as a viable testing ground for new technologies' that are unregulated and are 'deployed in opaque spaces', without sufficient international regulation and accountability.[22] What results is the constitution of humanitarian subjects as suitable test subjects who become subservient to the aim of making the technology safe for more 'valuable citizens'.[23] Further, new technologies often rely on the collection of sensitive data, which can give rise to new protection needs because the collection of such data creates demands that it be protected, which has ramifications for the safety of vulnerable populations.[24] For example, 'women at risk' would be particularly vulnerable if their data were revealed to permit identification or their stories were uploaded to insecure networks.[25]

If humanitarian tech is not designed with the input of the population affected, it can 'uphold western epistemological frameworks that understand complexities as "problems"' that ignore the politics of borders and risk further entrenching inequity and exclusion.[26] When technology is presented as a decision to 'save lives' and 'empower' individuals, subjects of that technology are 'constructed as design opportunities for the generosity of the elite, rather than as historical subjects with their own worldviews,

[17] Reem Talhouk and others, 'Involving Syrian Refugees in Design Research: Lessons Learnt from the Field' (Designing Interactive Systems Conference, San Diego, 23–28 June 2019).

[18] Savita Bailur and Bjorn-Soren Gigler, *Closing the Feedback Loop: Can Technology Bridge the Accountability Gap?* (World Bank 2014) 6.

[19] Everett Roger, *Diffusion of Innovations* (5th edn, New York Free Press 2003).

[20] Katja Lindskov Jacobsen and Larissa Fast, 'Rethinking Access: How Humanitarian Technology Governance Blurs Control and Care' (2019) 43 Disasters S151, S157.

[21] ibid S159.

[22] Petra Molnar, 'Technology on the Margins: AI and Global Migration Management from a Human Rights Perspective' (2019) 8 Cambridge International Law Journal 305, 305.

[23] Jacobsen and Fast (n 20) S156.

[24] Katja Lindskov Jacobsen and Kristin Bergtora Sandvik, 'UNHCR and the Pursuit of International Protection: Accountability through Technology?' (2018) 39 Third World Quarterly 1508, 1518.

[25] ibid 1516.

[26] Keshavarz (n 11) 31.

skills, and political sensibilities'.[27] Technology can result in the transfer of risk to the very populations it is intended to help. For example, technology designed to increase security and safety allows security personnel, especially those in senior positions, to operate remotely, which removes staff from 'the field' and potentially leaves local workers and populations exposed to attack.[28]

The potential of any new system overlaying or integrating with an existing system cannot be reduced to simplicities that measure whether the new system is 'better'. Viewing the use of AI in RSD as part of the digital transformation of the refugee regime[29] creates an impetus to understand how its opportunities and risks will affect both upward accountability (for example, decision-making efficiencies and more efficient and targeted aid delivery)[30] and downward accountability (that is, impacts on refugees). Taking a socio-legal perspective of AI in RSD means that before we ask how it may create positive outcomes for its subjects, we must consider what gaps it will need to fill. In other words, how well does the current credibility assessment system serve refugees?

3. THE 'HUMAN PROBLEM' OF CREDIBILITY ASSESSMENTS AND SUBJECTIVE FEAR

Credibility assessments, which are a crucial element[31] in the RSD process, aim to determine whether a claimant has a 'well-founded fear of being persecuted'.[32] In establishing whether an application is credible, the decision maker 'gathers information and examines it in light of all the information available, and then determines whether that information can be considered material to the application'.[33] In the context of RSD, material facts are based upon evidence such as the claimant's oral and written testimony, expert reports, witness testimony, and country of origin information (COI). Material facts are those that are 'credible, relevant and significant'[34] and are given weight by the decision

[27] ibid 24.

[28] Jori Pascal Kalkman, 'Practices and Consequences of Using Humanitarian Technologies in Volatile Aid Settings' (2018) 3 International Journal of Humanitarian Action 1.

[29] Kristin Bergtora Sandvik, 'The Digital Transformation of Refugee Governance' in Cathryn Costello, Michelle Foster, and Jane McAdam (eds), *The Oxford Handbook of International Refugee Law* (Oxford University Press 2021) 1008.

[30] Jacobsen and Sandvik (n 24) 1514.

[31] A significant proportion of decisions to deny status are based wholly or partially on adverse credibility findings: see UNHCR, 'Quality in the Swedish Asylum Procedure: A Study of the Swedish Migration Board's Examination of and Decisions on Applications for International Protection' (2011).

[32] Refugee Convention (n 4) art 1A(2).

[33] UNHCR and European Refugee Fund of the European Commission, *Beyond Proof: Credibility Assessment in EU Asylum Systems* (2013) 27 <https://www.unhcr.org/en-au/protection/operations/51a8a08a9/full-report-beyond-proof-credibility-assessment-eu-asylum-systems.html> accessed 24 September 2021.

[34] *Applicant VEAL of 2002 v Minister for Immigration and Multicultural and Indigenous Affairs* (2005) 225 CLR 88.

maker because they are convincing and have the potential to affect the outcome of the decision.[35]

Credibility assessments are complicated by the bipartite standard of a 'well-founded fear of being persecuted', which involves consideration of subjective fearfulness and the objective validation of that fear.[36] Although the threshold of 'truth' does not need to be met in credibility assessments, the decision maker faces the challenge of needing to be satisfied that the claimant can demonstrate an objective risk *and* subjective fear. National courts[37] and UNHCR have confirmed the bipartite standard. The latter has said:

> To the element of fear – a state of mind and a subjective condition – is added the qualification 'well-founded'. This implies that it is not only the frame of mind of the person concerned that determines his refugee status, but that this frame of mind must be supported by an objective situation. The term 'well-founded fear' therefore contains a subjective and an objective element, and in determining whether well-founded fear exists, both elements must be taken into consideration.[38]

It has been suggested that, lawfully and normatively, there is no subjective element in the 'well-founded fear' standard.[39] Whether a forward-looking apprehension of risk mandates a purely objective inquiry,[40] the subjective element imposes an additional burden on the refugee. A claimant may be found not to be a refugee if they are not perceived to be 'subjectively fearful', irrespective of evidence of an objective risk.[41] Documentary or other evidence may demonstrate a well-founded fear of being persecuted, regardless of what the claimant has said and whether the decision maker believes it.[42] The Federal Court of Canada expressed similar concerns in *Yusuf v Canada*:

35 Niamh Kinchin, 'Technology, Displaced? The Risks and Potential of Artificial Intelligence for Fair, Effective, and Efficient Refugee Status Determination' (2021) 37 Law in Context 45, 52. It is noted that this article by one of the authors is drawn upon throughout the present article. All references are provided.

36 James C Hathaway and Michelle Foster, *The Law of Refugee Status* (2nd edn, Cambridge University Press 2014) 91–92.

37 *Re Minister for Immigration and Multicultural Affairs, ex parte Miah* (2001) 206 CLR 57, 76 (Gaudron J); *Ward v Canada (Attorney General)* [1993] 2 SCR 689, 723 (La Forest, L'Heureux-Dubé, Gonthier, and Iacobucci JJ); *Zgnat'ev v Minister for Justice, Equality and Law Reform* [2001] IEHC 70, para 6 (Finnegan J).

38 UNHCR, *Handbook on Procedures and Criteria for Determining Refugee Status and Guidelines on International Protection under the 1951 Convention and the 1967 Protocol relating to the Status of Refugees*, HCR/1P/4/ENG/REV.4 (1979, reissued 2019) para 38.

39 See eg James C Hathaway and William S Hicks, 'Is There a Subjective Element in the Refugee Convention's Requirement of Well-Founded Fear?' (2005) 26 Michigan Journal of International Law 505; Michael Bossin and Laila Demirdache, 'A Canadian Perspective on the Subjective Component of the Bipartite Test for "Persecution": Time for Re-Evaluation' (2004) 1 Refuge 108.

40 Hathaway and Foster (n 36) 105.

41 *Bela v Canada* [2013] FC 784.

42 Michael Kagan, 'Is Truth in the Eye of the Beholder? Objective Credibility Assessment in Refugee Status Determination' (2003) 17 Georgetown Immigration Law Journal 367, 370.

> [But] I find it hard to see in what circumstances it could be said that a person … could be right in fearing persecution and still be rejected because it is said that fear does not actually exist in his conscience. The definition of a refugee is certainly not designed to exclude brave or simply stupid persons in favour of those who are more timid or more intelligent. Moreover, I am loath to believe that a refugee status claim could be dismissed solely on the ground that as the claimant was a young child or a person suffering from a mental disability, he or she was incapable of experiencing fear, the reasons for which clearly exist in objective terms.[43]

The requirement that claimants show evidence of objective *and* subjective fear gives rise to the 'human problem' of refugee credibility assessments. The decision maker must determine a claimant's plausibility, often without evidence beyond that person's own testimony.[44] There are no standard criteria to guide the decision maker, which necessitates subjective decision making that is 'highly personal to the decision-maker, dependent on personal judgment, perceptions, and disposition, and often lacking an articulated logic'.[45] Inevitably, there is an increased risk of bias and reviewable inaccuracies.

A need to guide decision makers in undertaking a structured inquiry[46] has prompted the development of 'credibility indicators'. Credibility indicators require decision makers to rely upon concrete evidence that reveals material facts and avoids speculative reasoning reflecting 'the decision-maker's personal theory of how the applicant could or should have acted, or how certain events could or should have unfolded'.[47] Varying according to jurisdiction, some examples of credibility indicators are those created by UNHCR and the European Union (EU) in their 2013 credibility assessment guide (as represented in figure 1).[48]

Despite their promise of objectivity, credibility indicators do not remove the risk of decision makers making speculative assessments of the inner workings of a claimant's mind. In requiring the decision maker to 'take into account such factors as the reasonableness of the facts alleged', which are 'on balance, capable of being believed',[49] credibility indicators adopt the conventions of an objective inquiry based on an examination of a claimant's actions and statements.[50] The objective language, whilst avoiding reliance on 'gut feelings' by encouraging the consideration of consistency, coherence, and corroborative evidence, can become what Hathaway and Foster call 'objective surrogate indicators of supposedly subjective fear'.[51]

43 *Yusuf v Canada* (1992) 1 FC 629, para 5 (Hugessen, Marceau, and MacGuigan JJA).

44 Kinchin (n 35) 54.

45 Kagan (n 42) 374.

46 ibid.

47 UNHCR and European Refugee Fund of the European Commission (n 33) 77.

48 ibid ch 5. See also UK Home Office, *Asylum Policy Instruction: Assessing Credibility and Refugee Status* (Version 9.0, 2015) 13.

49 UNHCR, 'Note on Burden and Standard of Proof in Refugee Claims' (1998) para 11 <https://www.refworld.org/docid/3ae6b3338.html> accessed 12 July 2021.

50 Bossin and Demirdache (n 39) 111.

51 Hathaway and Foster (n 36) 96–97.

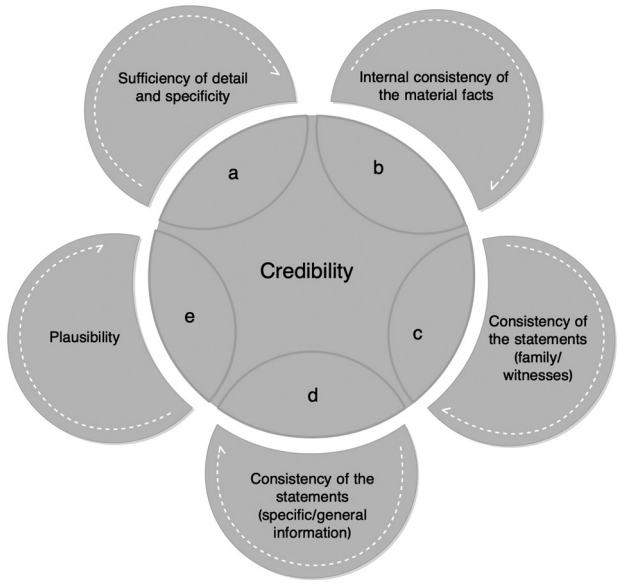

Figure 1. Credibility indicators in refugee status determination

See, for instance, the example below of a decision that was reviewed by UNHCR:

> Additionally one could ask the question why you haven't taken any initiative in X to verify what happened after you left Iraq. […] This nonchalant attitude towards your problems and your lack of initiative shows a lack of interest in your situation, which can hardly be explained given the circumstances under which you left your country. This puts your invoked fear for persecution in perspective. Of an asylum seeker it can be expected that he makes efforts to be informed about the reasons that pushed him to leave and that he informs himself about the evolution of his personal problems. That you neglected to take these actions casts serious doubts over your asylum claim and gives serious presumptions that you left Iraq for different reasons than you indicated.[52]

The objective indicators enable the decision maker to speculate about what an asylum seeker 'can be expected' to do, whilst ignoring any compounding factors that might have explained his 'lack of initiative' and 'lack of interest' in his situation. As Evans Cameron points out, information that is 'central' and 'peripheral' remains a subjective determination that is 'to be made from the perspective of the person whose memory is at issue', that is, the claimant.[53]

In assessing objective risk, the indicators provide a structured inquiry into the claimant's experience of past persecution, the persecution of others in comparable situations, and generalized risk.[54] At the same time, however, they continue to encourage

[52] UNHCR and European Refugee Fund of the European Commission (n 33) 208.

[53] Evans Cameron (n 3) 484.

[54] *Rajudeen v Canada (Minister of Employment and Immigration)* (1984) 55 NR 129, 134 (Heald J).

assessments of cognition, demeanour, and risk response that give rise to speculative reasoning. The requirement for evidence of subjective fear is an unnecessary burden for claimants who are already so burdened by circumstance, and is a complication for a decision-making process already characterized by uncertainty. The following discussion explores how credibility indicators, which have been folded into 'sufficiency of detail and specificity', 'consistency', and 'plausibility', perpetuate the difficulties of assessing subjective fearfulness as an indicator of refugee credibility.

3.1 Sufficiency of detail and specificity

A decision maker is required to consider a claimant's level of detail or knowledge about relevant events or circumstances. Regard should be given to matters that the claimant, taking into account their personal characteristics, within context, could reasonably be expected to know.[55]

The assessment of credibility on the basis of sufficient detail and specificity is challenged by the unreliability of memory. Memory is influenced by higher cognitive interactions with personality, mood, and the perceived intentions of the interviewer,[56] so vagueness, gaps, and inconsistencies in testimony[57] can indicate difficulties in retrieving and expressing memories, as well as with sufficiency of knowledge. The retrieval and expression of memory are also impacted by recognized difficulties in remembering dates and times,[58] the frequency of events,[59] names,[60] and everyday or common objects, such as money.[61]

When trauma is layered on top of issues of memory retrieval, the requirement for refugees to demonstrate sufficient knowledge and detail becomes more problematic. Disturbing events, even without trauma or mental health issues, can alter memory.[62] Increased arousal during a traumatic event can lead to concentration on specific detail with reduced recall of peripheral detail.[63]

[55] UNHCR and European Refugee Fund of the European Commission (n 33) 138; *Nejad v Minister for Immigration and Multicultural Affairs* [1999] FCA 1827, para 9 (Tamberlin J) upheld on appeal in *Nejad v Minister for Immigration and Multicultural Affairs* [2000] FCA 741; *Wang v Minister for Immigration and Multicultural Affairs* [2000] FCA 963, paras 20–23 (Goldberg J); *T v Minister for Immigration and Multicultural Affairs* [2000] FCA 467, paras 27–47 (Drummond, Matthews, and Mansfield JJ).

[56] Juliet Cohen, 'Questions of Credibility: Omissions, Discrepancies and Errors of Recall in the Testimony of Asylum Seekers' (2001) 13 International Journal of Refugee Law 293, 295.

[57] Hilary Evans Cameron, *Refugee Law's Fact-Finding Crisis* (Cambridge University Press 2018) 27.

[58] John S McIntyre and Fergus IM Craik, 'Adult Age Differences for Item and Source Information' (1987) 41 Canadian Journal of Psychology 175, cited in Cohen (n 56) 296.

[59] Gillian Cohen and Rosalind Java, 'Memory for Medical History: Accuracy of Recall' (1995) 9 Applied Cognitive Psychology 273, 274, cited in Evans Cameron (n 3) 476.

[60] Evans Cameron (n 3) 486.

[61] ibid 476.

[62] Evans Cameron (n 57) 48.

[63] Sven-Åke Christianson and others, 'Eye Fixations and Memory for Emotional Events' (1991) 17 Journal of Experimental Psychology, Learning, Memory and Cognition 693.

A 2014 study compared the ability of refugees with and without depression and post-traumatic stress disorder (PTSD), which are usually comorbid, to recall specific memories from their past. The researchers found that refugees with PTSD and depression were less able to provide examples of specific personal memories and were more likely to report extended memories,[64] which often reflected extended trauma, such as detention. Others failed to respond at all.[65] Rape and sexual assault victims with PTSD often demonstrate recall deficits. The decision maker needs to appreciate that 'the act of remembering a traumatic event does not necessarily equate to a coherent and consistent memory or narrative, as fragmentation can be part of the memory structure itself'.[66]

The way an interview is conducted, including the way questions are framed, can have an impact on recall. People tend to remember more with repeated recall ('hypermnesia'),[67] so the number of interviews, and the number of times a person is required to repeat their story, is significant. The questions asked in an interview are either free recall (that is, open questions that have no cues), or closed recall (that is, closed questions that have cues and give suggestions for the target information). An example of an open question used in refugee interviews would be: 'What happened to you after the military seized power?'[68] An example of a closed question would be: 'You said that you hid with your brother but on your basic data form you have indicated that your only brother lives abroad. How many brothers do you have?'[69] Although closed questions with cues may trigger more detailed recall than open questions,[70] cues may be misleading and provoke falsely remembered details. Open-ended questions may lead to distress.[71]

Other mitigating factors can cause refusal or inability to reveal specific and detailed knowledge. People are less likely to reveal things that are personally embarrassing[72] or where trust is still developing.[73] For example, some will find it shameful to discuss sexual orientation.[74] In some societies, fathers or other males may not have shared details of the claim with the females in the family, or will speak on their behalf. This may result in a female claimant providing only short or limited answers to the questions

64 Belinda Graham, Jane Herlihy, and Chris R Brewin, 'Overgeneral Memory in Asylum Seekers and Refugees' (2014) 45 Journal of Behavior Therapy and Experimental Psychiatry 375.

65 Extended memories are a type of overgeneral memory that reflects events that occurred repeatedly or lasted a long time; they can be contrasted with autobiographical memories, that relate to discrete events or short time periods, usually lasting less than one day.

66 Gillian McFadyen, 'Memory, Language and Silence: Barriers to Refuge within the British Asylum System' (2018) 17 Journal of Immigrant and Refugee Studies 168, 172.

67 David G Payne, 'Hypermnesia and Reminiscence in Recall: A Historical and Empirical Review' (1987) 101 Psychological Bulletin 5.

68 UNHCR, 'Interviewing Applicants for Refugee Status' (Training Module RLD4, 1995) 12.

69 ibid 14.

70 ibid 296.

71 Cohen (n 56) 300.

72 *Bibi v Immigration and Naturalization Service* 203 F 3d 830 (9th Cir 1999).

73 *SAAK v Minister for Immigration and Multicultural Affairs* [2002] FCAFC 86.

74 *WAIH v Minister for Immigration* [2003] FMCA 40, para 23 (Raphael FM).

posed.[75] For women, there can be shame in discussing sexual abuse and assault, since in some societies there is a stigma about loss of virginity, even when caused by rape. Women may be reluctant to share details about certain experiences with a male official or through a male interpreter.[76] A lack of education can also affect a claimant's capacity for self-expression.[77]

Insufficient detail in a claimant's account may reflect a lack of awareness of the relevance of specific details to an application. This may be due to inadequate space on an application form, a lack of guidance regarding the significance of specificity, or a misunderstanding of the questions,[78] which may not be designed to elicit appropriate detail.

Finally, claimants may simply be silent. Silence is often perceived as a sign of evasiveness and unresponsiveness and is 'subject to the imposition of unsolicited meaning'.[79] The United Kingdom (UK) Home Office states that 'failure without reasonable explanation to answer a question asked by a deciding authority' 'must always be treated as damaging to the claimant's credibility'.[80] However, as McFadyen points out, 'silences within a testimony can be sites of knowledge in themselves, providing substance to a story'.[81] Silence may be a means of protection,[82] and silent pauses can indicate traumatic events.

3.2 Consistency

According to the 'consistency heuristic', 'consistency implies truth, whereas inconsistency implies deception'.[83] Consistency can refer to consistency within the claimant's own evidence, consistency of the claimant's statements with information provided by family members and/or witnesses, or consistency with available specific and general information, such as COI.

Consistency must be 'material' and 'immediately relevant' to the claim.[84] According to the European Court of Human Rights, the credibility of statements and supporting documents should only be questioned where inconsistencies affect the core of the claimant's story.[85] Nonetheless, decision makers are often directed to draw a negative inference from inconsistencies. Take, for example, the UK Home Office, which said:

[75] Jane Herlihy, Laura Jobson, and Stuart Turner, 'Just Tell Us What Happened to You: Autobiographical Memory and Seeking Asylum' (2012) 26 Applied Cognitive Psychology 661.

[76] Refugee, Asylum and International Operations Directorate (RAIO), US Citizenship and Immigration Services, 'Asylum Officer Basic Training Module on Gender-Related Claims' (2012) para 7.1.2.

[77] ibid para 7.1.4.

[78] UNHCR and European Refugee Fund of the European Commission (n 33) 145.

[79] Toni AM Johnson, 'On Silence, Sexuality and Skeletons: Reconceptualizing Narrative in Asylum Hearings' (2011) 20 Social and Legal Studies 57, 57.

[80] UK Home Office, *Asylum Policy Instruction* (n 48) 39.

[81] McFadyen (n 66) 175.

[82] ibid 177.

[83] Leif A Strömwall, Pär A Granhag, and Anna-Carin Jonsson, 'Deception among Pairs: "Let's Say We Had Lunch and Hope They Swallow It!"' (2003) 9 Psychology, Crime and Law 109, 121, cited in Evans Cameron (n 3) 490.

[84] *Sabaratnam v Canada (Minister of Employment and Immigration)* [1992] FCJ No 901, para 1 (Mahoney JA).

[85] *FH v Sweden* App No 32621/06 (ECtHR, 20 January 2009) para 95.

> It is reasonable to expect that an applicant who has experienced an event will be able to recount the central elements in a broadly consistent manner. An applicant's inability to remain consistent throughout both written and oral accounts of past and current events may lead the decision maker not to believe the claim.[86]

Inconsistency may arise from a difference in phrasing, rather than from inconsistent facts. For instance, one refugee participant in a study on discrepancies in memory changed his description of his treatment by military police from 'we were slapped around' to 'we were badly beaten' when describing the same event.[87]

Memory also has implications for consistency because it can be reconstructed and reported differently each time.[88] Over time, memory can lose specificity and our 'older memories tend to be less vivid and less accessible', even though our confidence in our memories often remains misguidedly high.[89] It is common for claimants in RSD to rehearse memories in order to minimize natural inconsistencies. However, rehearsing some memories can come at the expense of others, and may make it harder to recall the types of details that are considered indicators of credibility.[90] Conversely, some people may *gain* memory through hypermnesia and reminiscence (that is, where the person adds or finds memories as they try to memorize them).

A 2001 study, which involved twice interviewing refugees of Kosovan and Bosnian backgrounds, found that discrepancies in testimonies were common, even when there was no foreseeable reason for the discrepancies. The researchers concluded that discrepancies – which were common in peripheral, rather than central, details and which were impacted by PTSD and the length of time between interviews – did not equal a lack of credibility.[91]

Findings of inconsistencies may be due to external factors, such as decision maker error, the interview conditions, or other cultural and social factors that can affect the claimant's testimony. In the case of *Mibanga v Secretary of State for the Home Department*, the Court of Appeal of England and Wales found that the adjudicator:

> artificially separated the medical evidence from the rest of the evidence and reached conclusions as to credibility without reference to that medical evidence; and then, no doubt inevitably on that premise, found that the medical evidence was of no assistance to her.[92]

[86] UK Home Office, 'Considering Asylum Claims and Assessing Credibility' (2012) 14 <https://www.refworld.org/pdfid/5449ffa84.pdf> accessed 2 April 2021.

[87] Jane Herlihy, Peter Scragg, and Stuart Turner, 'Discrepancies in Autobiographical Memories: Implications for the Assessment of Asylum Seekers' (2002) 324 British Medical Journal 324, 324.

[88] Evans Cameron (n 3) 491.

[89] Craig R Barclay and Henry M Wellman, 'Accuracies and Inaccuracies in Autobiographical Memories' (1986) 25 Journal of Memory and Language 93, 101, cited in Evans Cameron (n 3) 471.

[90] Evans Cameron (n 3) 494.

[91] Herlihy, Scragg, and Turner (n 87).

[92] *Mibanga v Secretary of State for the Home Department* [2005] EWCA Civ 367, para 30 (Buxton LJ).

When the retrieval method shifts between face-to-face interviews and self-administered questionnaires, memory can be impacted.[93] The act of remembering may be affected by the subject's impression of what is being asked and by the desire 'to please the interviewer by producing events about which they are being questioned'.[94] Compounding issues include cultural and gender factors, trauma, stress, fatigue, and language and interpretation issues.[95] Where interpreters are used, the potential for misunderstandings increases exponentially. If interpreters are also torture victims, or closely involved with such victims, they may close off certain questions or, where the interpreter is physically present, give non-verbal cues discouraging elaboration of detail.[96]

3.3 Plausibility

Plausibility determines whether the asylum seeker will be 'believed'. UK guidance states that plausibility of a fact is assessed on the basis of its 'apparent likelihood or truthfulness in the context of the general country information relevant to the claimants' country of origin and/or their own evidence'.[97] Although plausibility depends on the specificity, detail, and internal consistency of a claimant's testimony,[98] it may also be affected by the claimant's demeanour, as well as decision-maker assumptions and biases.[99]

Some commentators suggest that demeanour should have no place in credibility assessments,[100] while others merely caution against its use, without discrediting it.[101] The Canadian position, for example, is that 'the [Refugee Protection Division (RPD) of the Immigration and Refugee Board of Canada] can evaluate the general demeanour

93 Seymour Sudman and Norman M Bradburn, 'Effects of Time and Memory Factors on Response in Surveys' (1973) 68 Journal of the American Statistical Association 805, 815, cited in Evans Cameron (n 3) 507.

94 David C Rubin and Alan D Baddeley, 'Telescoping Is Not Time Compression: A Model of the Dating of Autobiographical Events' (1989) 17 Memory and Cognition 653, 658, cited in Evans Cameron (n 3) 496.

95 For the latter, see *Sherpa v Canada (Minister of Citizenship and Immigration)* [2009] FCJ No 665, paras 23–24, 57 (Russell J), in which the court found that an interpreter was 'sufficiently precise and competent to convey [the claimant's] words on the material points of concern', even though she had on several occasions mistranslated the Board's questions to the claimant, had 'inaccurately translated her answers and explanations, as well as adding words she had not said', had used English words in interpreting to the claimant on 270 occasions, and had 'acknowledged during the hearing that [the claimant] was having difficulty understanding her because they were from different localities and had different accents'.

96 Cohen (n 56) 300. See also *Zubeda v Ashcroft* 333 F 3d 463 (3rd Cir 2003) 476–77 (Circuit Judge McKee).

97 UK Border Agency, *Asylum Instructions: Considering Asylum Claims and Assessing Credibility* (2012) para 4.3.6.

98 Jane Herlihy, Kate Gleeson, and Stuart Turner, 'What Assumptions about Human Behaviour Underlie Asylum Judgments?' (2010) 22 International Journal of Refugee Law 351, 361.

99 Kinchin (n 35) 55.

100 Kagan (n 42) 380. See also UK Home Office, *Asylum Policy Instruction* (n 48) 18.

101 *SAAK v Minister for Immigration and Multicultural Affairs* [2002] FCA 367, para 27 (North, Goldberg, and Hely JJ).

of a witness' but that 'relying on demeanour to find a claimant not credible must be approached with a *great deal of caution*'.[102]

Many factors can affect demeanour – an individual's personality traits, age, gender, sexual orientation and/or gender identity, maturity, culture, social status, education, and psychological and physical states.[103] As the evaluation of demeanour involves 'assessing the manner in which the witness replies to questions, his or her facial expressions, tone of voice, physical movements, general integrity and intelligence, and powers of recollection',[104] such evaluations in cross-cultural settings are inherently problematic.[105] In one cultural perception study comparing Chinese with Western Caucasian participants, the Chinese participants were found to rely on eyes to represent facial expressions, whereas Western Causcasians relied on their eyebrows and mouths.[106] Another example of differing cultural perceptions relates to eye contact. In many non-Western cultures, the lowering of the eyes, particularly for women, is considered socially respectful. However, in Western asylum systems, this may be perceived as a sign of lying.[107]

People who have experienced trauma can show signs of depression, indecisiveness, indifference, poor concentration, long pauses before answering, avoidance, and disassociation.[108] The retelling of trauma may also elicit behaviour not usually associated with traumatic events, such as smiling or laughing nervously.[109]

Evaluation of demeanour is subjective and will 'inevitably reflect the views, prejudices, personal life experiences, and cultural norms of the decision-maker'.[110] 'Affinity bias' means that 'we tend to prefer people who look like us, think like us and come from backgrounds similar to ours'.[111] One study, which examined assumptions made by asylum decision makers, found that a claimant's taking action, rather than staying where they were, would be considered detrimental to their claim.[112] The following example of a decision was cited: 'I do consider it implausible that a family in fear, on seeing a man throw something over the fence and into their garden … would go to investigate it'.[113]

[102] Immigration and Refugee Board of Canada, *Assessment of Credibility in Claims for Refugee Protection* (2004) para 2.3.7 (emphasis added).

[103] ibid.

[104] ibid.

[105] See *Dia v Ashcroft* 353 F 3d 228 (3rd Cir 2003).

[106] Rachael E Jack, Roberto Caldara, and Philippe G Schyns, 'Internal Representations Reveal Cultural Diversity in Expectations of Facial Expressions of Emotion' (2011) 141 Journal of Experimental Psychology: General 19.

[107] Andrew P Bayliss and Steven P Tipper, 'Predictive Gaze Cues and Personality Judgments: Should Eye Trust You?' (2006) 17 Psychological Science 514, cited in McFadyen (n 66).

[108] Immigration and Naturalization Service, US Department of Justice, 'Guidelines for Children's Asylum Claims' (1998) 14.

[109] McFadyen (n 66) 174.

[110] UNHCR and European Refugee Fund of the European Commission (n 33) 186.

[111] Kimberly A Houser, 'Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making' (2019) 22 Stanford Technology Law Review 290, 321.

[112] Herlihy, Gleeson, and Turner (n 98) 358.

[113] ibid.

It was found that judges made assumptions that asylum seekers would know how to lodge an application and navigate the asylum process.[114] The Court of Appeal of England and Wales has warned that, in determining that an account is not credible, the decision maker must take care not to do so merely because what is described would be implausible had it occurred in the UK,[115] and that 'decision-makers must constantly be on guard to avoid re-characterizing the nature of the risk based on their own perceptions of reasonability'.[116]

People who are perceived to act differently from those in the same group or minority may be treated with suspicion. LGBTQI+ people, for example, can find that they are forced to challenge assumptions about their behaviours and actions because RSD encodes and reflects homophobic stereotyping.[117] In one example, the court thought the claimant's mannerisms were not 'gay enough'.[118] There is also a tendency to make assumptions about how women should act, even though vulnerability may explain behaviour such as staying in an abusive relationship. Children's testimony must also be treated carefully as they can appear uncooperative and their testimony may seem ambiguous.[119]

A 'culture of disbelief' is created[120] when decision makers employ their own mental protective devices to resist the negative effects of hearing about upsetting events[121] or are influenced by a shared institutional culture through 'institutionally sanctioned documents', such as official guides or policy, which are largely based on generalizations about the behaviour and motives of particular social groups.[122]

4. FROM A 'HUMAN PROBLEM' TO A DATA PROBLEM, OR AN OPPORTUNITY FOR REFORM?

The integration of AI into RSD may provide invaluable decision-making support if it can address the flaws inherent in human assessments of credibility. An examination of the potential form and function of AI in RSD, and an analysis of its risks and opportunities, reveals that AI's potential for enhancing upward and downward accountability will not be realized if subjective fear remains a determinant of a claimant's credibility.

114 ibid.

115 *Y v Secretary of State* [2006] EWCA Civ 1223.

116 *HK v Secretary of State for the Home Department* [2006] EWCA Civ 1037, para 29 (Neuberger LJ), quoting James C Hathaway, *The Law of Refugee Status* (1st edn, Butterworths 1991) 81.

117 Catherine Dauvergne and Jenni Millbank, 'Burdened by Proof: How the Australian Refugee Review Tribunal Has Failed Lesbian and Gay Asylum Seekers' (2003) 31 Federal Law Review 299.

118 *Todorovic v Attorney General* 6221 F 3d 1318 (11th Cir 2010). See also *Raza v Minister for Immigration and Multicultural Affairs* [2002] FCAFC 82.

119 *Abay v Ashcroft* 368 F 3d 634 (6th Cir 2004).

120 McFadyen (n 66) 177.

121 Richard F Mollica, 'The Trauma Story: The Psychiatric Care of Refugee Survivors of Violence and Torture' in Frank M Ochberg (ed), *Post Traumatic Therapy and Victims of Violence* (Brunner/Mazel 1988), cited in Cohen (n 56).

122 Smith-Khan (n 3) 418, 424.

4.1 The use of artificial intelligence in refugee status determination: form and function

Fundamental to the design and operability of many AI systems is machine learning. Machine learning, which applies algorithms that 'learn' from historic data to identify correlations, can be developed using both unsupervised or supervised learning techniques. Unsupervised learning is based upon pre-programmed logic where rules are coded into a system and data matching is used to draw inferences about future behaviour based on 'if (condition), then (action)' logic. Unsupervised learning does not autonomously learn, but relies upon input by human experts. For example, components of a decision-making process that rely on 'clear, fixed and finite criteria', such as legislative criteria for determining eligibility for a benefit, would be considered unsupervised learning.[123]

Supervised learning is where algorithms autonomously learn from training or historic data to identify correlations, patterns, or clusters so that input or new data can be given as a probability, which can be used to inform a decision.[124] Predictions, or probabilities, are based on inductive or deductive reasoning, which involves the 'transformation of data to knowledge by making inferences, planning and scheduling activities, searching through a large solution set, and optimising solutions'.[125] In automating the construction of the rules that drive the system, input data must be pre-classified or labelled, based on the patterns and correlations the machine has 'learnt' from the training data, or set.

Machine learning could support the RSD process through 'expert systems'. 'Expert systems' utilize unsupervised and supervised learning techniques, depending on the nature of the problem and the decisions they need to make.

A data-mining system that analyses COI by creating clusters and discovering patterns based on different topics or hierarchies is an expert system that could assist decision makers to identify relevant information and to corroborate the claimant's testimony. A comparable system already exists in the form of e-discovery. This is used by law firms to replicate and replace traditionally labour-intensive discovery tasks by observing how lawyers review documents. E-discovery 'learns' the criteria that make a document relevant until the system reaches a threshold of confidence to prioritize documents for review[126] and inclusion in litigation. If a training dataset could be created from COI in a

[123] Monika Zalnieriute, Lyria Bennett Moses, and George Williams, 'The Rule of Law and Automation of Government Decision-Making' (2019) 82 Modern Law Review 8, 9.

[124] Hilary Evans Cameron, Avi Goldfarb, and Leah Morris, 'Artificial Intelligence for a Reduction of False Denials in Refugee Claims' (2021) 34 Journal of Refugee Studies 3.

[125] High-Level Expert Group on Artificial Intelligence, 'A Definition of AI: Main Capabilities and Scientific Disciplines' (2018) 4 <http://www.pcci.gr/evepimages/0101_F483.pdf> accessed 22 January 2021.

[126] Ajith Samuel, 'Artificial Intelligence Will Change E-Discovery in the Next Three Years' (*Law Technology Today*, 24 April 2019) <https://www.lawtechnologytoday.org/2019/04/artificial-intelligence-will-change-e-discovery-in-the-next-three-years/> accessed 12 January 2020.

similar way, then key-value pairs[127] might be extracted from the claimant's forms, travel and identity documentation, and testimony. Key-value pairs could be compared to the COI training dataset to identify patterns that could help corroborate the claimant's evidence.

A form of expert system already in use in the refugee context is 'biometrics'.[128] 'Biometrics' refers to techniques that measure a person's physical characteristics to verify their identity through physiological traits,[129] such as fingerprints, facial features, irises, and even DNA, voice verification, vein pattern recognition, and keystroke dynamics.[130] Since 2013, UNHCR has operated the Biometric Identity Management System (BIMS), which captures and stores fingerprints, iris data, and facial images. BIMS is currently being used in 60 operations, with 7.2 million biometric records collected up to 2019.[131] Eighty per cent of UNHCR refugees over the age of five now have a biometric identity.[132] Biometrics can be used in the registration and reception stages of RSD to help establish a claimant's identity and to provide the evidentiary link between an asylum seeker and their testimony. The nexus between identity and credibility is acknowledged in the Immigration and Refugee Protection Act 2001 (Canada):

> The Refugee Protection Division must take into account, *with respect to the credibility* of a claimant, whether the claimant possesses acceptable documentation establishing identity, and if not, whether they have provided a reasonable explanation for the lack of documentation or have taken reasonable steps to obtain the documentation.[133]

A third potential expert system that might be used in RSD is automated credibility assessments. Technology-enhanced credibility assessments are traditionally associated with polygraph screening or lie detection tests. While early machines used perspiration and heart rate to measure whether a person was lying, today's deception detection

[127] As Cheng explains, 'Key-Value Pairs are two linked data items, a key, and a value, where the key is used as a unique identifier for the value': see Tim Cheng, 'How to Extract Key-Value Pairs from Documents Using Deep Learning' (*Nanonets*) <https://nanonets.com/blog/key-value-pair-extraction-from-documents-using-ocr-and-deep-learning/> accessed 28 December 2020.

[128] Lina Chato and Shahram Latifi, 'Application of Machine Learning to Biometric Systems: A Survey' (2018) 1098 Journal of Physics: Conference Series <https://iopscience.iop.org/article/10.1088/1742-6596/1098/1/012017> accessed 27 September 2021.

[129] Nikolas Kairinos, 'The Integration of Biometrics and AI' (2019) Biometric Technology Today 8.

[130] Lindsey N Kingston, 'Biometric Identification, Displacement, and Protection Gaps' in Carleen F Maitland (ed), *Digital Lifeline? ICTS for Refugees and Displaced Persons* (MIT Press 2018).

[131] UNHCR, 'Introductory Remarks of Andrew Harper: Director of the Division of Programme Support and Management', UN doc EC/70/SC/CRP.13 (19 June 2019).

[132] 'Data of Millions of Refugees Now Securely Hosted in PRIMES' (*UNHCR Blogs*, 28 January 2019) <https://www.unhcr.org/blogs/data-millions-refugees-securely-hosted-primes/> accessed 1 April 2020.

[133] Immigration and Refugee Protection Act 2001 (Canada) s 106 (emphasis added).

includes measuring a person's eye movements or 'gaze behaviour', 'vocalics',[134] kinesics and proxemics,[135] linguistic analysis,[136] and cardiorespiratory and thermal measures.[137] Such credibility assessments have been tested in pilot projects in border control and security. From 2016 to 2019, Hungary, Latvia, and Greece trialled an automated lie detection test called 'iBorderCtrl', which questioned international travellers and used AI to record and analyse facial micro-gestures in order to determine whether the travellers were telling the truth. Before arriving at the airport, the traveller was required to log on to a website and upload an image of their passport. Once the image was uploaded, the person was asked a number of questions about themselves and the purpose of their travel by an avatar. Using the webcam, the program scanned the traveller's face and eye movements for perceived signs of lying. At the end of the interview, the system issued a QR code that was shown to a border guard. The guard scanned the code, took fingerprints, and reviewed the facial image to check that it corresponded with the traveller's passport. The guard's device then displayed a score out of 100, indicating whether the system had judged the traveller to be truthful, or not.[138]

Similarly, the Automated Virtual Agent Truth Assessment in Real Time (AVATAR), a system developed by the United States (US) National Center for Border Security and Immigration (BORDERS), uses a virtual agent to conduct credibility interviews while 'simultaneously detecting potential anomalous behaviour via analysis of data streams from non-invasive sensors such as cameras, microphones, and eye-tracking systems'.[139]

A more nuanced form of automated credibility assessments exists in the field of human resources, where automated interviews determine a person's 'appropriateness' for a role or task. Sensor devices are used to analyse facial expressions, body language, and gestures, as well as the emotional sentiment of voice and text, to determine whether a person is 'a

[134] Some of the vocal cues important to the automated detection of deception are voice quality, pitch, and response latency. These three measures are reliable indicators of either the stress or cognitive load associated with deception: see Jay F Nunamaker and others, 'Establishing a Foundation for Automated Human Credibility Screening' (Information Systems and Quantitative Analysis Faculty Proceedings and Presentations, University of Nebraska at Omaha, 2012) 3.

[135] These measure lip presses, chin raises, fidgeting, facial pleasantness, lack of movement or rigidity, pupil dilation, blinking patterns, eye movements, and overall tension ratings to differentiate truth from deception: ibid.

[136] In research on automated credibility assessments of *intentions*, Kleinberg and others integrated 'verbal deception theory' with computer-automated analysis to detect truthful and deceptive statements about planned activities using the Linguistic Inquiry and Word Count (LIWC) program. LIWC reads a given text and counts the percentage of words that reflect different emotions, thinking styles, and social concerns. See Bennett Kleinberg and others, 'Automated Verbal Credibility Assessment of Intentions: The Model Statement Technique and Predictive Modeling' (2018) 32 Applied Cognitive Psychology 354.

[137] Nunamaker and others (n 134).

[138] Ryan Gallagher and Ludovica Jona, 'We Tested Europe's New Lie Detector for Travelers and Immediately Triggered a False Positive' (*The Intercept*, 26 July 2019) <https://theintercept.com/2019/07/26/europe-border-control-ai-lie-detector/> accessed 8 April 2021.

[139] Jay F Nunamaker and others, 'Field Tests of an AVATAR Interviewing System for Trusted Traveler Applicants' (2013) <https://eller.arizona.edu/sites/default/files/FieldTestsofanAVATARInterviewingSystemforTrustedTravelerApplicants.pdf> accessed 7 January 2020.

good fit' for the company. HireVue[140] is designed to determine whether a candidate will be friendly, empathetic, and meet the employer's requirements by parsing 'videos using machine learning, extracting signals like facial expression and eye contact, vocal indications of enthusiasm, word choice, word complexity, topics discussed, and word groupings'.[141] The program 'uses these signals to create a model that claims to capture relationships between interview responses and workplace performance, based on the employer's pre-existing metrics'.[142]

Intelligent digital forensics assist the decision maker in 'identification, acquisition, preservation, analysis and presentation' of digital documentation[143] through the automation of classification and processing to extract key-value pairs and entities, clustering documents into different topics or hierarchies, and analysing and tagging files by extracting metadata.[144] Digital documentation in the context of RSD may include digitized versions of official documents, as well as digital data in the possession of asylum seekers. Social network and mobile phone data possessed by the asylum seeker – such as contacts, short message service (SMS) messages, instant messaging text and media, location records, and browsing history[145] – could be digitally analysed to help establish the identity of claimants, corroborate facts where inconsistencies in testimony arise, and help fill in gaps where there is insufficient detail.[146]

Finally, machine learning may be designed to support human RSD decision makers through algorithms that are based on supervised learning. Signifying an evolution from earlier 'non-AI' decision support systems in refugee decision making,[147] algorithms

[140] 'HireVue Hiring Platform: Video Interviews, Assessment, Scheduling, AI, Chatbot' (*HireVue*) <https://www.hirevue.com/> accessed 25 August 2022.

[141] Amanda Rogen and Aaron Rieke, 'Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias' (*Analysis and Policy Observatory*, 9 December 2018) 35 <https://apo.org.au/node/210071> accessed 15 January 2020.

[142] ibid.

[143] Stefania Costantini, Giovanni DeGasperis, and Raffaele Olivieri, 'Digital Forensics and Investigations Meet Artificial Intelligence' (2019) 86 Annals of Mathematics and Artificial Intelligence 193, 194.

[144] David Parmenter, 'The State of AI in Document Management' (*Adobe Blog*, 22 October 2019) <https://theblog.adobe.com/state-of-ai-in-document-management/> accessed 12 January 2020.

[145] Morgan Meaker, 'Europe Is Using Smartphone Data as a Weapon to Deport Refugees' (*Wired*, 2 July 2018) <https://www.wired.co.uk/article/europe-immigration-refugees-smartphone-metadata-deportations> accessed 13 January 2020.

[146] Existing research uses the communication and mobility (real-world call detail record) of refugees to measure the integration of Syrian refugees in Turkey: see Antonio Luca Alfeo and others, 'Assessing Refugees' Integration via Spatio-Temporal Similarities of Mobility and Calling Behaviors' (2019) 6 IEEE Transactions on Computational Social Systems <https://arxiv.org/ftp/arxiv/papers/1907/1907.06929.pdf> accessed 25 August 2022.

[147] John Yearwood, 'Case-Based Retrieval of Refugee Review Tribunal Text Cases' (JURIX: The Tenth Conference, Amsterdam, 12 December 1997) 67; John Yearwood and Andrew Stranieri, 'The Integration of Retrieval, Reasoning and Drafting for Refugee Law: A Third Generation Legal Knowledge Based System' (Proceedings of the 7th International Conference on Artificial Intelligence and Law, Oslo, 14–17 June 1999) 117; John Zeleznikow, 'Building Decision Support Systems in Discretionary Legal Domains' (2000) 14 International Review of Law, Computers and Technology 341.

would learn from training data to help predict whether an individual has a well-founded fear of being persecuted. The training data would include data on past cases,[148] or previous RSD outcomes, which could be clustered according to 'incident type' (for example, violent attacks on, or denial of services to, individuals belonging to particular socio-cultural groups) and jurisdiction, country of origin, and asylum country.

The input data in RSD, which would involve information available about a particular case, would be used to generate predictions of the likelihood of an event or a response occurring. For example, 'in order to generate a prediction on the likelihood that a particular police force responds adequately to a particular type of appeal for help, the input data would be all the information available about the specific appeal for help and police force in a given case'.[149] This information could include official information, such as COI and government documentation. COI reports 'collate relevant information on conditions in countries of origin pertinent to the assessment of claims for international protection'[150] and provide up-to-date insight into the conditions that cause displacement. Government documentation may include travel records, identity documents, and expert testimonials. Available information may also include documentation that claimants themselves possess, such as identity and travel documents and written testimonials. The primary type of information drawn upon for credibility assessments, however, is likely to be claimants' oral testimony.

4.2 Opportunities for the application of artificial intelligence in refugee status determination

4.2.1 Fairness testing and removing bias

If, as discussed below, AI systems carry and reflect unconscious bias – which, in turn, can impact accuracy and create discrimination – they may also be able to identify and remove it.

Algorithmic bias can be addressed *prior* to the implementation of an expert or decision-making system. Methods include creating balanced datasets – which involves 'boosting' a less frequent category, such as traditionally 'female' traits – and increasing the diversity of data points so that more data is reviewed than is usually measured. Although credibility assessments *distinguish* characteristics such as race, gender, and country of origin, 'similar cases should be treated alike and result in the same outcome' in RSD.[151] 'Unbiased algorithms' could guide the decision maker to distinguish physical, social, and cultural traits to take them into account, rather than remove them. Algorithmic design could also be required to be more transparent, and the selection of programmers could be required to reflect greater demographic diversity.[152]

148 Evans Cameron, Goldfarb, and Morris (n 124) 13.

149 ibid.

150 UNHCR, 'Refugee Status Determination' <https://www.unhcr.org/en-au/refugee-status-determination.html#:~:text=Country%20of%20Origin%20Information%20(COI)%20is%20information%20which%20is%20used,of%20claims%20for%20international%20protection> accessed 24 December 2020.

151 Note that the EU also adopts this approach: see European Council, 'The Stockholm Programme: An Open and Secure Europe Serving and Protecting Citizens' [2010] OJ C115/01, 32.

152 Houser (n 111) 336–39.

A possible option for retrospective bias identification is 'fairness testing'. Fairness testing uses 'fairness metrics' to detect algorithmic bias and discrimination *after* a system is designed, but before it becomes operational. For example, Themis,[153] a system developed by researchers from the University of Massachusetts, detects the cause of discriminatory behaviour. If an algorithm is to be fair, the same output must be produced for every two individuals who differ only in respect of relevant characteristics. The software measures the fraction of inputs for which changing other, non-relevant characteristics causes the output to change. In RSD, the algorithm might identify that gender is the characteristic that causes the output (or decision) to change for two people with otherwise similar claims.

4.2.2 Removing 'noise'

AI systems might address 'noise' in RSD. 'Noise' refers to variability in human decision making due to chance or irrelevant factors.[154] Noise in RSD could refer to the kinds of questions asked, the use of interpreters, or decision-maker error. Although they are part of the decision-making process, such factors should not influence substantive outcomes. In a US case, *Santashbekov v Lynch*, the Court of Appeals, Seventh Circuit, was concerned that asylum seekers were being prompted to provide different levels of detail in their testimony, depending on the design of the form:

> The I-589 Asylum Application Form provides small boxes to detail an applicant's experiences, containing space for about ten lines of text. We caution against drawing adverse credibility conclusions from an applicant providing differing levels of detail in such different contexts. The limited space on the I-589 form provides a readily apparent reason why [the applicant] was able to provide a more detailed account of his alleged persecution at the hearing than on the application.[155]

Creating rules that are consistently applied to datasets can reduce noise and create consistency, or even uniformity.[156] Automated credibility interviews could ensure that the nature and environment of a credibility interview are consistent. Standardizing questions and the formats of interviews would help improve consistency because different retrieval cues (that is, questions) generate varying recollections.[157]

Some research suggests that interviewees may actually respond better to virtual environments.[158] A 2018 criminal justice study found that where the interviewer

[153] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou, 'Fairness Testing: Testing Software for Discrimination' (Proceedings of the 11th Joint Meeting on Foundations of Software Engineering, Paderborn, 4–8 September 2017) 498.

[154] Houser (n 111) 323.

[155] *Santashbekov v Lynch* 834 F 3d 836 (7th Cir 2016) para 7 (Circuit Judge Hamilton).

[156] Houser (n 111) 331.

[157] Julian AE Gilbert and Ronald P Fisher, 'The Effects of Varied Retrieval Cues on Reminiscence in Eyewitness Memory' (2006) 20 Applied Cognitive Psychology 723, cited by Evans Cameron (n 3) 506.

[158] But see Liz Bradley and Hillary B Farber, 'Virtually Incredible: Rethinking Deference to Demeanor when Assessing Credibility in Asylum Cases Conducted by Video Teleconference' (2022) 36 Georgetown Immigration Law Journal 515. Bradley and Farber's study highlights the challenges faced by asylum seekers involved in virtual interviews.

and interviewee were represented by avatars, episodic recall improved, compared to face-to-face interviews. The researchers hypothesize that the demand characteristics[159] associated with the physical presence of the interviewer may have been reduced, which decreased errors arising from a real or perceived pressure to perform. The study found that, as interviewees did not have to engage with the situational dynamics of the interview context, more cognitive resources were available to facilitate episodic retrieval.[160]

Technological challenges remain. For example, linguistic analysis based on automated transcription requires calibration to an individual's voice before an acceptable level of accuracy can be obtained. Calibration takes time, which may preclude its application to rapid screening applications[161] and undermine efficiency. Further, the setting for an automated interview must be free from distractions. A noisy environment, whether in a person's home or in another location, could affect sensors, responses, and interactions.[162] These are not insurmountable challenges, however, and could be factored into the design process.

4.2.3 Fact-finding and consistent, up-to-date country of origin information

Machine learning that creates clusters and discovers patterns in COI databases to identify relevant information and corroborate the claimant's testimony could enhance the fact-finding process. Although the burden of proof in RSD rests with the claimant, the decision maker is bound by a shared duty of fact-finding. According to UNHCR, 'in some cases, it may be for the examiner to use all the means at his disposal to produce the necessary evidence in support of the application.'[163] Machine learning and expert systems could help recognize out-of-date, incorrect, or even biased sources of information, especially where inconsistencies arise. The capacity of expert systems to locate and process large amounts of data could assist the decision maker where doubts are raised about the consistency of the claimant's testimony, or where there are gaps in the details that the person has provided.

Humanitarian data analytics could help to create better COI information. Satellite imagery and information about economic, political, geographic, and meteorological circumstances,[164] including human rights violations and ethnic and civil conflicts, could contribute to the development of up-to-date, consistent COI databases.[165]

[159] Subtle cues that let participants know that they are expected to answer or act in a given way.

[160] Donna A Taylor and Coral J Dando, 'Eyewitness Memory in Face-to-Face and Immersive Avatar-to-Avatar Contexts' (2018) 9 Frontiers in Psychology 507.

[161] Nunamaker and others (n 134).

[162] ibid.

[163] UNHCR Handbook (n 38) para 196.

[164] Susan Martin and Lisa Singh, 'Data Analytics and Displacement: Using Big Data to Forecast Mass Movements of People' in Carleen F Maitland (ed), *Digital Lifeline? ICTS for Refugees and Displaced Persons* (MIT Press 2018) 185, 190, 197.

[165] Kinchin (n 35) 59.

4.2.4 Revealing uncertainty

It may be that AI in RSD will have the capacity to make clear to decision makers how uncertain their predictions are. Exposing the lack of available information explicitly forces decision makers to confront the uncertainty inherent in RSD.[166] Resolving doubt in legal decision making requires an application of the burden of proof. As Evans Cameron, Goldfarb, and Morris argue, if the law resolved decision-making doubt in favour of refugee protection, AI could increase the number of claims accepted, which should lead to a reduction in the number of rejected claims.[167] A change in the law to shift the burden from the refugee, or at least a more explicit and generous policy of applying the benefit of the doubt,[168] would be required if AI were to have such an impact.

Alternatively, AI's ability to identify irrelevant factors or characteristics may act as a support tool for human decision making to *reduce* uncertainty.

4.3 Risks in the application of artificial intelligence in refugee status determination

4.3.1 Replicating the 'human problem' of credibility assessments

AI might replicate the human problem in refugee credibility assessments, but with less capacity for resolution. The difficulties in assessing a claimant's plausibility are somewhat ameliorated by a human decision maker's capacity for critical self-reflexivity. Critical self-reflexivity is the process of questioning one's own assumptions, presuppositions, and meaning perspectives.[169] Self-reflection allows greater sensitization to intercultural communication[170] and other factors, such as memory and trauma, that can have an impact upon a claimant's testimony. When supported through appropriate training or mentoring, critical self-reflexivity can allow decision makers to recognize and acknowledge their unconscious bias and the mental shortcuts they use to process information.[171] Although humans may not always choose to be critically self-reflective, they retain the capacity to be so.

An algorithm, however, cannot learn self-reflexivity if human biases and assumptions are embedded in pre-labelled data.[172] Two examples provide important reminders of the risks of incorporating unconscious bias into a system. The first, a study that compared Face++ and Microsoft AI (two facial recognition services), analysed a dataset of photos of professional basketball players for emotional perception. Algorithmic bias was evident in both services. Face++ categorized results based on the 'level of smile', and interpreted black players as angrier for every level. Microsoft AI interpreted

166 Evans Cameron, Goldfarb, and Morris (n 124).

167 ibid 18.

168 Smith-Khan (n 3) 426.

169 John Mezirow, 'An Overview of Transformative Learning' in Peter Sutherland and Jim Crowther (eds), *Lifelong Learning: Concepts and Contexts* (Routledge 2006) 24–38.

170 Audrey Macklin, 'Truth and Consequences: Credibility Determination in the Refugee Context' (International Association of Refugee Law Judges 3rd Conference, Ottawa, 14–16 October 1998) 134, cited by Smith-Khan (n 3) 426.

171 Houser (n 111) 173.

172 Zalnieriute, Moses, and Williams (n 123) 8.

black players as more contemptuous than non-black players in ambiguous and/or non-smiling pictures.[173] The second example concerned a 2017 recruitment exercise by Amazon, which used an AI-based recruitment system that vetted applicants by observing patterns in resumes submitted to the company over a 10-year period. The exercise had to be abandoned because the system 'learnt' to disregard activities that were associated with women, such as team sports with 'women' in the name, because most of the resumes had come from men.[174] Algorithms rely upon 'good' pre-labelling of data to avoid incorporating unconscious bias into a system.[175]

Similar concerns arise when algorithms, like humans, are designed to make assumptions based on an institutional narrative. Smith-Khan argues that, although an RSD decision-making process may appear to be individualized in a way that 'can obviate the impacts of social and linguistic diversity in order to evaluate objectively and justly the credibility of asylum-seekers and their claims',[176] decisions inevitably reflect an institution's narrative. An institutional narrative – manifested through 'institutional texts[, namely] credibility assessment guidelines and a set of published "merits review" appeals decisions' – represents how a decision-making body perceives the main social actors involved in RSD.[177] For example, asylum seekers are represented as 'applicants', meaning that their institutional function of 'applying' for a visa overshadows broader human rights protections. Alternatively, the decision maker is represented as the reasonable institutional insider and the receiver of information.[178] Whereas the human decision maker internalizes an institutional narrative in a way that allows the possibility of critical self-reflexivity, an algorithm would embed that narrative in a way that does not allow the same possibility.

Complicating the absence of algorithmic self-reflexivity is the fact that the communication of emotions presents particular challenges for AI. The expression of emotions often affects how credible a claimant is deemed to be. The way in which people communicate emotions such as sadness and disgust varies across cultures and situations, and even similar configurations of facial movements can express more than one type of emotion.[179] 'Emotion recognition algorithms' could include a protocol designed to control, or at least minimize, the effects of moderating factors such as culture, context, question type, personality, and situational factors.[180] However, algorithms are unlikely

[173] Lauren Rhue, 'Racial Influence on Automated Perceptions of Emotions' (2018) <https://ssrn.com/abstract=3281765> accessed 30 September 2021.

[174] Jeffrey Dastin, 'Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women' (*Reuters*, 11 December 2018) <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> accessed 21 April 2020.

[175] Houser (n 111) 333.

[176] Smith-Khan (n 3) 424.

[177] ibid 407.

[178] ibid 412–13.

[179] Lisa Feldman Barrett and others, 'Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements' (2019) 20 Psychological Science in the Public Interest 1.

[180] Nunamaker and others (n 134).

to be better at assessing emotions because they work best when their purpose is to look at content and behaviour from a skills test, rather than how people sound or look. Further, the data inefficiency of algorithms means that they 'typically need millions of examples to learn distinctions that would strike a human as immediately obvious'.[181] Machine learning would need an unfeasible quantity of data in order for an algorithm to 'learn' when the way a claimant expresses emotions and recalls memories may be impacted by trauma and other cultural, social, psychological, or physical issues.

Any technology that has the potential to affect the rights and obligations of vulnerable populations must be considered within a human rights framework.[182] The prospect of algorithms replicating the human problem in credibility assessments by manifesting biases and assumptions threatens equality rights and freedom from discrimination.[183] For example, algorithms that learn based on COI data that ignores or diminishes threats to the safety of some groups in that society, such as the LGBTQI+ community,[184] will increase the risk of discrimination.

4.3.2 The prospective nature of a well-founded fear

The prospective nature of a well-founded fear may compromise the ability of AI to act as a support tool in decision making. Traditional legal processes are unsuited to reliable identification of genuine trepidation because historical evidentiary facts 'do not provide a sound basis for a determination that any asylum seeker is entitled to protection *now*'.[185] Evidence of past persecution acts as a 'guide' as to what may happen if a person returns to their country of origin,[186] but assessment of risk is speculative.[187] The fact that threats have not been carried out does not render a person's fear unreasonable. What is important is whether the 'group making the threat has the will and ability to carry it out'.[188]

[181] David Watson, 'The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence' (2019) 29 Minds and Machines 417, 423.

[182] Molnar (n 22) 319.

[183] International Human Rights Program, Faculty of Law, University of Toronto, and Citizen Lab, Munk School of Global Affairs and Public Policy, University of Toronto, 'Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System' (2018) 30 <https://ihrp.law.utoronto.ca/sites/default/files/media/IHRP-Automated-Systems-Report-Web.pdf> accessed 25 July 2020. Equality and anti-discrimination rights are embedded in a number of international instruments, an example being the International Convention on the Elimination of All Forms of Racial Discrimination (adopted 21 December 1965, entered into force 4 January 1969) 660 UNTS 195.

[184] International Human Rights Program and Citizen Lab (n 183) 33.

[185] *R (Saber) v Secretary of State for the Home Department* [2007] UKHL 57, para 2 (Lord Hope) (emphasis added).

[186] *S152/2003 v Minister for Immigration and Multicultural Affairs* [2004] HCA 18, para 74 (McHugh J). See also *Katrinak v Secretary of State for the Home Department* [2001] EWCA Civ 832.

[187] Guy S Goodwin-Gill and Jane McAdam, *The Refugee in International Law* (4th edn, Oxford University Press 2021) 56–62.

[188] *Marcos v Gonzales* 410 F 3d 1112 (9th Cir 2005) para 16 (Circuit Judge Paez).

The forward-looking nature of a well-founded fear means that decision makers cannot rely upon experience alone to improve their decision making. They can rarely verify whether their decisions were correct or incorrect, or on what grounds.[189] Without such feedback, decision makers are confronted by uncertainty, which they must resolve by identifying and avoiding what Evans Cameron terms the 'wrong mistake', or the 'worst' outcome.[190] Indeed, predictions of well-founded fear could really only be 'accurate' 'when applied to an environment that is sufficiently regular to be predictable, and when, through prolonged exposure, there is an opportunity to recognize its regularities'.[191]

If the requirement for a prospective assessment of potential harm poses challenges for the human decision maker, it creates new ones for machine learning. Algorithms rely on input data, which is labelled according to the patterns and correlations that the machine has 'learnt' from the training data. It follows that the training data, which is historical, can only be labelled according to evidence of *past* persecution in countries or regions, based on the outcomes of other cases. Whilst positive findings of persecution in similar situations may assist the decision maker to identify whether the claimant has an objectively provable fear of being persecuted, past persecution cannot provide a 'ground truth' or objectively provable data on which the algorithm can base its predictions. It will also provide no insight into subjective fear. It follows, as Evans Cameron, Goldfarb, and Morris argue, that 'given the sparse data and uncertain environment in refugee claims, any machine predictions in this context are likely to generate wide distributions'.[192]

The findings of a 2019 study that used machine learning to predict the outcome of judgments from the European Court of Human Rights serve to highlight the unique challenges faced by the application of machine learning in RSD.[193] The study utilized a natural language-processing program to analyse the texts of judgments (that is, the training data) in order to predict whether any article of the European Convention on Human Rights (ECHR)[194] was violated. An algorithm called a support vector machine (SVM) linear classifier was applied to automatically predict the category (that is, a verdict of violation or non-violation) associated with a new element (that is, a case, or input data). Unlike RSD, the study could rely on past evidence of violations, rather than a speculation as to whether violations *would be* committed. Whilst providing an insight into how predictive algorithms may be used in litigation, the study also serves to spotlight the unique obstacles posed by a (risk of) prospective harm in the 'well-founded fear' standard.

189 Evans Cameron (n 3) 509.

190 Evans Cameron (n 57) 9.

191 Daniel Kahneman, *Thinking Fast and Slow* (Allen Lane 2011) 259–60.

192 Evans Cameron, Goldfarb, and Morris (n 124) 2.

193 Masha Medvedeva, Michel Vols, and Martijn Wieling, 'Using Machine Learning to Predict Decisions of the European Court of Human Rights' (2020) 28 Artificial Intelligence and Law 237.

194 Convention for the Protection of Human Rights and Fundamental Freedoms (adopted 4 November 1950, entered into force 3 September 1953) 213 UNTS 222 (European Convention on Human Rights) (ECHR).

A closer analogy to RSD may exist in algorithmic predictions of future offending within the criminal justice system, which are known as 'risk assessments'. Risk assessments statistically analyse court and police records to identify which factors, such as the number of prior convictions and a person's employment, housing, or education status, might best predict recidivism.[195] Despite their prospective nature, risk assessments fail to provide useful insights into a well-founded fear because they predict the risk of future events based on past events, rather than the 'risk of risk', which is a more attenuated concept. In RSD, claimants may 'prevail on a theory of future persecution despite an … adverse credibility ruling as to past persecution so long as the factual predicate of [their] claim of future persecution is independent of the testimony that the [immigration judge] found not to be credible'.[196]

Humans address the challenges presented by prospectivity through abductive reasoning, which is where the decision maker accepts one of multiple competing hypotheses based on relative plausibility.[197] Although abductive reasoning is a part of machine learning,[198] a suggestion that 'machine-based abduction' should, or could, be applied to competing hypotheses of a well-founded fear is neither ethically sound, nor feasible. The vulnerable nature of refugees and the significant impact of RSD on individuals' lives and access to opportunities speak against leaving such 'high-stakes decision-making processes' to AI alone.[199] Feasibility is compromised by the difficulties machine learning would have in resolving 'radical uncertainty'[200] in RSD. Human decision makers decide what 'version of truth' to accept by allocating an 'error burden', which is represented by the legal constructs of standard of proof and burden of proof.[201] Although standards of proof in refugee law vary across jurisdictions, such as a 'real chance' in Australia,[202] 'reasonable possibility' in the US,[203] 'reasonable degree of likelihood' in the UK,[204] and 'reasonable chance', reframed as 'serious possibility', in Canada,[205] all accept some level of speculation. Machine learning cannot speculate in a vacuum. As there will be insufficient or unreliable training data in RSD,[206] the lack of

[195] Megan T Stevenson and Jennifer L Doleac, 'Algorithmic Risk Assessment in the Hands of Humans' (2019) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440> accessed 26 August 2022.

[196] *Boika v Holder* 727 F 3d 735 (7th Cir 2013) para 13 (Circuit Judge Hamilton).

[197] Abductive reasoning is also known as 'inference to the best explanation'. See Gilbert H Harman, 'The Inference to the Best Explanation' (1965) 74 Philosophical Review 88.

[198] Zhi-Hua Zhou, 'Abductive Learning: Towards Bridging Machine Learning and Logical Reasoning' (2019) 62 Science China Information Sciences <http://scis.scichina.com/en/2019/076101.pdf> accessed 26 August 2022.

[199] Sarah Myers West, Meredith Whittaker, and Kate Crawford, *Discriminating Systems: Gender, Race, and Power in AI* (*AI Now Institute*, 2019) 6 <https://ainowinstitute.org/discriminatingsystems.pdf> accessed 24 April 2020.

[200] Luker (n 3) 515.

[201] Evans Cameron (n 57) 7–8.

[202] *Chan v Minister for Immigration and Ethnic Affairs* (1989) 169 CLR 379.

[203] *Immigration and Naturalization Service v Cardoza-Fonseca* 467 US 407 (1987).

[204] *R v Secretary of State for the Home Department, ex parte Sivakumaran* [1988] 1 All ER 193.

[205] *Adjei v Canada (Minister of Employment and Immigration)* [1989] 2 FC 680.

[206] Evans Cameron (n 57) 12.

structured inputs from which to extract patterns means that machine learning systems are unlikely to be able to make a determination about refugee status that has no apparent relation to its prior knowledge.[207]

4.3.3 Shifting the burden of proof to the refugee

'Digital fact-finding' may create a larger burden on the claimant to demonstrate a well-founded fear of being persecuted. Although digital identification and categorization of information could support a decision maker by providing a more efficient and perhaps more accurate means of fact-finding, the inevitable use of that technology to corroborate evidence is problematic. It is generally accepted that corroboration is not required where the testimony offered is credible, persuasive, and specific.[208] Only corroborating material that is 'readily available or material facts that are central' to the claim should be provided.[209] Corroboration should not be routinely insisted upon and should only be sought 'where such corroboration is needed to enable the decision-maker meaningfully to assess risk'.[210] The EU Qualification Directive[211] allows the governments of Member States to decide whether applicants for refugee status will be expected to corroborate oral testimony,[212] and there are no adverse implications if they choose not to.

A system designed to corroborate evidence may create an incremental creep where corroboration becomes an *expectation*, regardless of the credibility of the claim. The human decision maker must be left with the discretion to determine whether the claimant's 'testimony is credible, is persuasive, and refers to specific facts sufficient to demonstrate that the applicant is a refugee'[213] before deciding to corroborate the facts. The decision maker should not rely on a system that automatically considers COI as 'a test of direct corroboration, rather than for the purposes of providing context and general credibility assessment'.[214]

4.3.4 Deception, not credibility

Current automated credibility assessments largely rely on deception detection techniques, which look for 'truth' rather than 'credibility'. However, the role of credibility

[207] Zayan Guedim, 'Machine Learning vs Machine Reasoning: Know the Difference' (*Edgy*, 10 February 2019) <https://edgy.app/machine-learning-vs-machine-reasoning-know-the-difference> accessed 8 April 2021.

[208] *Re Dass* (1989) 20 I&N Dec 120, 124–25 (Chairman Milhollan, Board Members Dunne, Morris, Vacca, and Heilman).

[209] *Diallo v Immigration and Naturalization Service* 232 F 3d 279 (2nd Cir 2000) 288 (Chief Judge Walker).

[210] Hathaway and Foster (n 36) 156.

[211] Directive 2011/95/EU of 13 December 2011 on standards for the qualification of third-country nationals or stateless persons as beneficiaries of international protection, for a uniform status for refugees or for persons eligible for subsidiary protection, and for the content of the protection granted (recast) [2011] OJ L337/9 (EU Qualification Directive).

[212] ibid art 4(5).

[213] 8 USC §1158(b)(1)(B)(ii).

[214] *El Moraghy v Ashcroft* 331 F 3d 195 (1st Cir 2003) 204 (Circuit Judge Lynch).

assessments is not to determine the accuracy of a claimant's testimony. Indeed, a decision maker who is not certain of the veracity of a statement concerning a relevant fact may still find it credible and accept it for the purposes of RSD.[215] As Sweeney says, 'to show that a statement is credible is not the same as to show that it is true'.[216]

Even if deception techniques could be drawn upon, 'credibility cues' are linked to cognitive and behavioural reactions to real-world theoretical constructs. Accordingly, it would be important to replicate those theoretical constructs realistically, or the results will be unreliable.[217] Replicating real-world theoretical constructs in RSD is not plausible but, without them, false positives are likely. In August 2019, a journalist from the *Intercept* took part in the iBorderCtrl pilot scheme, described earlier, and returned a false positive. In the view of Ray Bull, professor of criminal investigation at the University of Derby, the project lacked credibility because there is no evidence that monitoring facial micro-gestures is an accurate way to measure whether people are telling the truth:

> They are deceiving themselves into thinking it will ever be substantially effective and they are wasting a lot of money … The technology is based on a fundamental misunderstanding of what humans do when being truthful and deceptive.[218]

4.3.5 Data abuse

Technologies that collect, mine, and analyse personal data, such as biometrics and the forensic analysis of mobile metadata and social media, pose risks to privacy rights and data protection.[219] As the Office of the UN High Commissioner for Human Rights points out, secret surveillance and 'the mere generation and collection of data relating to a person's identity, family or life' interfere with the right to privacy because the individual loses control of information that could put their privacy at risk.[220] The public sharing of information does not render it unprotected[221] and social media can exacerbate vulnerability when it is used as a tool for harm through, for example, 'deep fakes', disseminations of sexual activity, and false attribution of imagery.[222]

[215] UNHCR, 'Note on Burden and Standard of Proof in Refugee Claims' (n 49) 12: 'Given that in refugee claims, there is no necessity for the applicant to prove all facts to such a standard that the adjudicator is fully convinced that all factual assertions are true, there would normally be an element of doubt in the mind of the adjudicator as regards the facts asserted by the applicant'.

[216] James Sweeney, 'Credibility, Proof and Refugee Law' (2009) 21 International Journal of Refugee Law 700, 719.

[217] Nunamaker and others (n 134).

[218] Gallagher and Jona (n 138).

[219] International Human Rights Program and Citizen Lab (n 183) 40. For examples of relevant international human rights instruments, see eg International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171, art 17.

[220] UN Human Rights Council, *The Right to Privacy in the Digital Age: Report of the Office of the United Nations High Commissioner for Human Rights*, UN doc A/HRC/27/37 (30 June 2014) para 20.

[221] ibid para 6.

[222] Sandvik (n 29) 1021.

Biometrics poses particular risks for the abuse of data, partly because it is 'premised on a duty of refugee visibility'.[223] Identity theft based on biometrics is difficult to remedy[224] and biometric data may be used for purposes other than those for which the data were collected, including the unlawful tracking and monitoring of individuals.[225] International protection could be compromised by an absence of strong data protection legislation in refugee host countries if biometric data belonging to refugees is shared with either the host country or the country of persecution.[226] The misuse of biometric data may leave refugees exposed to discrimination and rights abuses if authoritarian States utilize biometric data to identify individuals and groups whose loyalty they question, and target them for surveillance or punitive action.[227]

As Sandvik and Jacobsen argue, the need to produce 'good data' contributes to the framing of biometrics and other forms of data collection as an accountability solution that results in less emphasis on downward accountability and more on upward accountability.[228]

5. 'WELL-FOUNDED FEAR' AS AN OBJECTIVE STANDARD

Algorithms are not better than humans at identifying subjective fear. Indeed, without the capacity for critical self-reflexivity, they may be worse. If data carry the unconscious biases and assumptions of the human developer, the machine may end up replicating and manifesting these when assessing the credibility of an asylum seeker. The forward-looking perspective of a well-founded fear, whilst supporting an objective inquiry, limits the ability of algorithms to predict persecution. Algorithms learn by utilizing training data that are historical and that can only be labelled according to evidence of past persecution in other places, based on the outcomes of other people's cases. Algorithms may prove useful for making predictions based on generalized risk, or even the previous outcomes of asylum adjudications,[229] but will be less effective in relation to the claimant's own, inevitably subjective, experiences of persecution and risk, and the experiences of those close to them.

If determination of a 'well-founded fear of being persecuted' were based on an objective standard only, the ability of AI to achieve greater standardization, and its capacity to mine and parse large amounts of data, would hold significant potential for increased consistency, improved fact-finding, and corroboration. Provided care were taken to ensure that corroboration did not become an expectation in every case, AI could help

223 ibid 1008.

224 *The Right to Privacy in the Digital Age* (n 220) 14.

225 ibid.

226 Kinchin (n 35) 60.

227 Jeff Crisp, 'Beware the Notion that Better Data Lead to Better Outcomes for Refugees and Migrants' (*Chatham House*, 9 March 2018) <https://www.chathamhouse.org/expert/comment/beware-notion-better-data-lead-better-outcomes-refugees-and-migrants> accessed 26 July 2021.

228 Jacobsen and Sandvik (n 24) 1514.

229 Daniel Chen and Jess Eagel, 'Can Machine Learning Help Predict the Outcome of Asylum Adjudications?' (Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law, London, 12–16 June 2017).

to shift the burden of proof to the decision maker. AI's capacity for fact-finding and corroboration would inevitably increase an expectation that the decision maker would take on a more inquisitorial role. With the benefit of more accurate and up-to-date facts, the decision maker would be forced to critically explore inconsistency or plausibility before arriving at a decision.[230]

AI will only benefit refugees if it does not replicate the problems of the current system. The integration of AI with RSD in a way that creates an effective and ethical humanitarian tech will depend on whether the 'well-founded fear of being persecuted' standard continues to be based on subjective fear *and* objective risk.

6. CONCLUSION

The integration and adaptation of AI into diverse areas of regulation and professional practice does not detract from its status as humanitarian tech when the populations it impacts are vulnerable. If AI in RSD is to be an ethical techno-human system, the harm that it risks creating for refugees, as well as its potential benefits, must be recognized. Assessments of harm and benefit cannot be disentangled from the challenges AI is being tasked to address. If AI reveals the failings of an existing system but cannot effectively address them, this does not necessarily mean that it is the technology that is weak. The possibility of utilizing AI to assist the RSD process shines a light on the flaws of current credibility assessments and the 'well-founded fear' standard.

Neither procedure nor legal principle is an obstacle to the removal of the subjective element from the 'well-founded fear' standard. Subjective fear is already disregarded in some situations, such as *prima facie* determinations. *Prima facie* determinations involve the removal of subjective fear in large-scale influx situations, recognizing refugee status on the basis of 'readily apparent, objective circumstances in the country of origin'.[231] Other RSD strategies that do not require the subjective element include accelerated case processing, enhanced registration processes, and simplified procedures. Alternatives to RSD, such as temporary protection arrangements and the suspension of RSD processing until the situation in a country becomes stable, also result in the removal of the subjective element.[232]

In this time of an 'AI spring', it is time to reconsider whether the requirement for asylum seekers to 'prove' subjective fear can continue to be justified.

[230] Rosemary Byrne, 'Assessing Testimonial Evidence in Asylum Proceedings: Guiding Standards from the International Criminal Tribunals' (2007) 19 International Journal of Refugee Law 609; Guy Coffey, 'The Credibility of Credibility Evidence at the Refugee Review Tribunal' (2003) 15 International Journal of Refugee Law 377.

[231] UNHCR, 'Guidelines on International Protection No 11: Prima Facie Recognition of Refugee Status', HCR/GIP/15/11 (24 June 2015) para 1.

[232] UNHCR Executive Committee, 'Refugee Status Determination', UN doc EC/67/SC/CRP.12 (31 May 2016) para 9.