Institution: University of Bath		
Unit of Assessment: B11 Computer Science and Informatics		
Title of case study: Artificial Intelligence & Ethics: Influencing Standards & Decision-Making and Remedying Bias		
Period when the underpinning research was undertaken: 2007-2020		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Joanna Bryson	Reader and Senior Research Fellow, previously Lecturer	October 2002 – September 2019
Period when the claimed impact occurred: 2014 – 2020		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact

Research at the University of Bath in Artificial Intelligence has impacted tools and design techniques, public policy, and industry practice in the understanding of ethics and mitigating unintended algorithmic bias. This research has:

- **Influenced UK, European and International standards in robot ethics**, including British Standards for Robots and Robotic Devices and standards for the Institute of Electrical and Electronics Engineers (IEEE).
- **Improved the understanding of policymakers** and influenced decision-making globally (USA, Canada, Europe, UK) through direct application of Bath research, advisory and expert roles on key committees (e.g., UK Government All-Party Parliamentary Group, International Committee of the Red Cross).
- **Reduced gender bias in Google Translate**, which, drawing directly on examples from Bath research, now provides feminine and masculine translations for some gender-neutral words.

2. Underpinning research

As Artificial Intelligence (AI) has transitioned from model-based approaches to increasingly data-driven machine learning techniques, resulting technologies have become more prone to various forms of bias, raising serious concerns over the ethical implications of AI and increasing demand from designers and policy makers to understand, mitigate and legislate for these biases.

Research by Bryson at the University of Bath has focused on exploring and addressing the ethical implications of AI and bias and, together with colleagues, developed ethical principles for robotics.

As AI techniques increasingly innovated in data-driven approaches, some in the field presumed that model-based work was of decreasing importance, whereas Bryson (with collaborators at Princeton University) recognised that the increasing lack of legibility would emphasise further the study of accountability in AI ethics and bias. This culminated in work which demonstrated that the use of natural language training corpora would result in models that embody cultural biases in a variety of forms, including gender and race. The authors demonstrated that tools such as Google Translate embed bias, for example when translating from a language without gendered pronouns to one with them. The research highlighted how Google Translate translates from Turkish's gender-free pronouns to 'he is a doctor' but 'she is a nurse' [**REF1**].

In further research (2017), Bryson and colleagues across the globe considered how we can guide the way technology impacts society. The methodologies that underpin these demonstrations have led to the development of new ethical principles, new standards and

standardization processes to ensure the safety, security, and reliability of AI [**REF2, REF3, REF4**]. Bryson argues that while making AI moral agents or patients is an intentional and avoidable action, avoidance would be our most ethical choice. Bryson argued (2010) that the potential of robotics should be understood as the potential to extend our own abilities and to address our own goals [**REF5**]. However, robots should not be described as persons, nor given legal nor moral responsibility for their actions [**REF2**]. Robots should also not have a deceptive appearance - they should not fool people into thinking they are similar to empathy-deserving moral patients. Bryson also argues that clear, generally-comprehensible descriptions of an artefact's goals and intelligence should be available to any owner, operator, or other concerned party [**REF4**]. Finally, Bryson's work has shown that the transparency of machine learning can be radically improved by providing real-time visualisation of a robot's AI, an approach that also helps an observer to understand the robot's behaviour [**REF6**].

3. References to the research

[REF 1] Caliskan, A, Bryson, JJ & Narayanan, A 2017, 'Semantics derived automatically from language corpora contain human-like biases', *Science*, vol. 356, no. 6334, pp. 183-186. https://doi.org/10.1126/science.aal4230

[REF 2] Bryson, JJ & Winfield, A 2017, 'Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems', *Computer*, vol. 50, no. 5, 7924235, pp. 116 - 119. https://doi.org/10.1109/MC.2017.154

[REF 3] Boden, M, Bryson, J, Caldwell, D, Dautenhahn, K, Edwards, L, Kember, S, Newman, P, Parry, V, Pegman, G, Rodden, T, Sorrell, T, Wallis, M, Whitby, B & Winfield, A 2017, 'Principles of robotics: regulating robots in the real world', *Connection Science*, vol. 29, no. 2, pp. 124-129. https://doi.org/10.1080/09540091.2016.1271400

[REF 4] Bryson, JJ 2018, 'Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics', *Ethics and Information Technology*, vol. 20, no. 1, pp. 15-26. https://doi.org/10.1007/s10676-018-9448-6

[**REF 5**] Bryson, JJ 2010, Robots should be slaves. in Y Wilks (ed.), *Close engagements with artificial companions: key social, psychological, ethical and design issues.* Natural Language Processing, vol. 8, John Benjamins Publishing Company, Amsterdam, pp. 63-74. https://doi.org/10.1075/nlp.8.11bry

[**REF 6]** Wortham, RH, Theodorou, A & Bryson, JJ 2017, Improving robot transparency: real-time visualisation of robot AI substantially improves understanding in naive observers. in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).*, 8172491, IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), vol. 26, IEEE, IEEE RO-MAN 2017, Lisbon, Portugal, 28/08/17. https://doi.org/10.1109/ROMAN.2017.8172491

4. Details of the impact

1) Influencing standards in robot ethics in the UK and worldwide

Bath research has directly influenced the development and content of the British Standard BS8611. BS8611 is the "*earliest explicit ethical standard in robotics*" and the "*one standard [that] specifically addresses AI*", providing "*guidance on how designers can identify potential ethical harm, undertake an ethical risk assessment of their robot or AI, and mitigate any ethical risks identified*" [**A,** p.4, 66]. "*At the heart of BS8611 is a set of 20 distinct ethical hazards and risks […] Advice on measures to mitigate the impact of each risk is given alongside suggestions on how such measures might be verified or validated*" [**REF4**]. The drafting of the standard resulted from Bryson's invitation to the UK Robot Ethics Forum (London, 2015); she provided further consultation on the development of BS8611 during 2015 with 5 of the 9 ethical principles set out

in 5.1.1. coming directly from Bryson's research [**REF4**]; the standard, BS8611:2016, is available to purchase via the BSI website [**K**].

The Institute of Electrical and Electronics Engineers (IEEE) is the world's largest technical professional organisation representing more than 400,000 members in over 160 countries. The University of Bath research influenced the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems and the IEEE P7000 series of standards, providing engineers and technologists with an implementable process to minimize ethical risk for their organizations, stakeholders and end users [**B**]. Bryson's work, cited in the IEEE manifesto *Ethically Aligned Design* [**B**], informed principles around transparency, embedding values into autonomous intelligent systems and Affective Computing. Since 2016 Bryson has co-chaired the IEEE Affective Computing Committee [**B**, p.14] and sits on the Sustainable Development group, that informs the P7000 standards.

2) Improving the understanding of policymakers

Bath research has led to improved understanding, mitigation and legislation for AI biases among designers and policymakers. This influence was achieved through Bryson's expert advice to UK, pan-European, and international governmental, inter-governmental and professional bodies.

UK Government Policy: Drawing on her research, Bryson influenced UK AI policy making, "*advising the Government to enable the on-demand and routine auditing of AI and algorithmic systems*" through her appointment as Expert Advisor to the All-Party Parliamentary Group on Artificial Intelligence (2017), in which she "*discussed the issue of algorithmic biases*" [**L**] and through other Expert Panel memberships [e.g., **C**]. Former Deputy Prime Minister Sir Nick Clegg stated "*Your presentation to the group (Open Reason Round Table Event in November 2017) and your contributions to the following discussion, was immensely useful…the round table played an integral role in preparing for the speech on the politics around artificial intelligence which I delivered at the end of last year (2017)*" [**D**].

Pan-European Policy: Bryson's research [**REFS 5, 6**] is cited in the European Parliamentary Research Service (EPRS) study on the ethical implications and moral questions arising from the development and implementation of AI technologies, including wealth inequality and political upheaval that could result from the rise in AI and the immorality of giving robots moral agency [**A**, p. 12, 14, 20, 35]. Further, the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation (CEN-CENELEC) cites Bryson's work in its Roadmap [**REF3; E**, p.29, 34] and notes that her research "*has been vital in making the case for AI standards in Europe*" [**J**].

International Representation: Among a number of influential roles, including: UN Centre for Policy Research (AI legal standards), Canadian Institute for Advanced Research, Mindfire ((Switzerland) Ethics Board); Bryson's research has informed the International Committee of the Red Cross (ICRC) around ethical issues raised by autonomous weapon systems and the requirement for human control over the use of force [**F**, p.1]. This work emphasised that states must urgently establish limits on autonomy in weapon systems and was cited in the ICRC 2020 report [**G**], commended "*primarily to government decision makers in the realms of international law, arms control, defence and foreign affairs*". This also led to the publication of an Accountability and Transparency report [**F**, p.4].

Bryson's influence is widely recognised: in 2017 Bryson was ranked as one of the Top 50 female artificial intelligence influencers in the world (Onalytica); and in 2019 Bryson was listed by Siliconrepublic as one of 10 AI influencers you should be following on Twitter [**H**].

3) Reducing gender bias in Google Translate

Bryson's research [**REF3**] informed changes in Google Translate operations. The Product Manager at Google Translate quotes Bryson's example word for word on Turkish gender-free pronouns [**REF3**] and explains the changes Google made: "*Our latest development in this effort*

addresses gender bias by providing feminine and masculine translations for some gender-neutral words on the Google Translate website … Now you'll get both a feminine and masculine translation for a single word […] For example, if you type 'o bir doktor' in Turkish, you'll now get 'she is a doctor' and 'he is a doctor' as the gender-specific translations" [**I**]. Google Translate is the world's most-used machine translation system, translating more than 100,000,000,000 words a day for 500,000,000 users.

5. Sources to corroborate the impact

[**A**] European Parliament. March 2020. "The ethics of artificial intelligence: Issues and initiatives". Panel for the future of science and technology. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

[**B**] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, IEEE, 2018. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

[**C**] Beard, Simon. APPG for Future Generations event: How do we make AI safe for humans? 20 July 2018. https://www.cser.ac.uk/news/appg-ai-safe/

[**D**] Letter from Former Deputy Prime Minister, 16 January 2018.

[**E**] CEN-CENELEC. "Focus Group on Artificial Intelligence (AI): Final draft for commenting: Roadmap from Focus Group on AI", 11 July 2020.

[**F**] Invited participant, "Ethics and autonomous weapon systems: An ethical basis for human control?" A roundtable panel of International Committee of the Red Cross (ICRC) "Accountability and Transparency." Humanitarium, Geneva, Switzerland, 28-29 August 2017. Contributed to ICRC report based on that meeting: Ethics and autonomous weapon systems: An ethical basis for human control? 3 April 2018. https://www.icrc.org/en/download/file/69961/icrc_ethics_and_autonomous_weapon_systems_report_3_april_2018.pdf

[**G**] Boulanin, V., Davison, N., Goussac, N., and Carlsson, M. P. 2020. Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control. https://www.icrc.org/en/document/limits-autonomous-weapons

[**H**] Influence awards: Dunmore, L. Women in Tech: Hot Topics and Top Influencers, 2017. https://onalytica.com/blog/posts/women-tech-hot-topics-top-influencers/; Darmody, J. 10 AI influencers you should be following on Twitter, 2019 https://www.siliconrepublic.com/people/ai-influencers-twitter

[**I**] James Kuczmarski, "Reducing gender bias in Google Translate", Google Translate Blog Post, December 6th 2018. https://www.blog.google/products/translate/reducing-gender-bias-google-translate/

[**J**] Testimonial Letter from Convenor of CEN-CENELEC AI Focus Group, 22 December 2020.

[**K**] BS 8611:2016. Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems, April 2016.

[**L**] Governance, Social and Organisational Perspective on AI. 11 September 2017. http://appg-ai.org/wp-content/uploads/2017/12/appgai_theme-report-5.pdf