Institution: University of Oxford

Unit of Assessment: 18 - Law

Title of case study: Increasing Transparency and Accountability in Algorithmic Decision Making

Period when the underpinning research was undertaken: March 2017 – December 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:
Prof. Sandra Wachter	Associate Professor and Senior Research Fellow	22 February 2017 – present

Period when the claimed impact occurred: March 2017 – December 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact (indicative maximum 100 words)

Wachter's research on the legal and ethical underpinning of AI systems has informed policy guidance as well as business practices in the UK, Europe and globally on algorithmic transparency, accountability and fairness as well as on data protection issues. Her work has impacted global policies and reports most notably the European Union's Article 29 Working Party's 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation', which closed several loopholes in the General Data Protection Regulation she previously identified. Following on from this work she developed tools to make AI human understandable and less biased. The research has provided practical tools to businesses, regulators, civil society and the judiciary to respond to both the policy and public pressure for greater algorithmic transparency and accountability. Google and Amazon are among the companies who have adopted Wachter's work on 'counterfactual explanations' as well as her bias test approaches, with implications for millions of people around the world.

2. Underpinning research (indicative maximum 500 words)

Sandra Wachter investigates the legal and ethical underpinnings of the complex AI systems that increasingly make decisions about our lives, despite varying degrees of human oversight and little explanation given to those affected. Her work on algorithmic accountability, explainability, and fairness has shed light on significant gaps in legal protections in the UK and Europe—but has also shown a practical way to make "black box" decisions human understandable, without compromising commercial interests.

In 2016 a persistent rumour arose among researchers, media and governments that the EU's General Data Protection Regulation (GDPR) would ban all machine learning systems that could not fully explain their logic when making decisions. If true, this requirement could have had unprecedented economic consequences for the European market. However, in **R1,** which Wachter started work on when she was at the Alan Turing Institute, but revised, edited, and published following her appointment at Oxford, working with Mittelstadt and Floridi, (University of Oxford, Unit of Assessment 30 – Philosophy), Wachter showed through an analysis of the legal framework and European case law that the GDPR does *not* guarantee such a blanket right to explanation. In fact—and contrary to prior research—she demonstrated that individuals lack meaningful protection against automated algorithmic decision-making, and she shed light on major loopholes in the framework that allowed industry to avoid such explanation if they chose.

While technical tools are being developed for computing explanations of algorithmic decisions, in **R2** Wachter shows that few provide a "good everyday explanation". Working with co-authors Mittelstadt (University of Oxford, ethicist) and Russell (University of Surrey, machine learning) she argues that people prefer contrastive and narrative explanations over technical decision

trees. That is, rather than being given technical explanations of the functioning of the underlying code, people prefer simple explanations of the most important factors in the decision, and how they would need to change to arrive at a different decision, for example: 'you were denied a loan because your income is GBP30,000; were your income GBP45,000 it would have been approved'.

Wachter next developed a method to provide such "counterfactual" explanations that are easily understood by the public, and easily generated by complex systems, while still protecting commercial interests. Combining law, ethics and computer science, **R3** was one of the first concrete and technically feasible solutions to compute "good everyday explanations" of decisions made by black box models, which are often described as fundamentally incomprehensible.

However, a clear decision is not necessarily a good one. In **R4** Wachter assesses the ethical legitimacy of the inferences and predictions on which companies base their automated decisions on things like hiring, credit, parole and insurance. She challenges the long-held idea that inferred data (e.g. on ethnicity) enjoys the same protection as other types of personal data in data protection law. In **R5** Wachter went on to ask whether EU non-discrimination law is equipped to deal with AI that infers sensitive information about individuals in order to target or exclude certain groups. She explains that the current law does *not* protect against these novel types of discrimination, and proposes practical ways to close these gaps.

Finally, **R6** analyses non-discrimination law and jurisprudence of the European Court of Justice (ECJ) and national courts, to identify an incompatibility between legal (i.e. contextual and intuitive) notions of discrimination, and standard technical measures of algorithmic fairness. Wachter shows that "automating fairness" in Europe may be impossible because the law does not provide a static framework for testing for discrimination in AI systems. Instead, the paper proposes 'conditional demographic disparity' (CDD) as a statistical measure of automated discrimination that aligns with the ECJ's gold standard for assessment of prima facie discrimination. This allows considerations of fairness to be built into automated systems while respecting the contextual approach to judicial interpretation practiced under EU non-discrimination law.

Wachter's research has been supported by a number of major grants, including funding from Engineering and Physical Sciences Research Council (EPSRC), DeepMind Technologies Limited, British Academy, Luminate Group, Alan Turing Institute, and the AI Fund of the Miami Foundation.

3. References to the research (indicative maximum of six references)

[R1] (Journal article) Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76-99. https://doi.org/10.1093/idpl/ipx005

[R2] (Conference contribution) Mittelstadt, B., Russell, C., & Wachter S. (2019). Explaining Explanations in AI. FAT* proceedings, 279-288. https://doi.org/10.1145/3287560.3287574

[R3] (Journal article) Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law and Technology, 31(2), 841-887 https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

[R4] (Journal article) Wachter, S. & Mittelstadt, B. (2019). A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI. Columbia Business Law Review 2019(2). https://doi.org/10.7916/cblr.v2019i2.3424

[R5] (Journal article) Wachter, S. (2020) Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. Berkeley Technology Law Journal 35 (2). http://dx.doi.org/10.2139/ssrn.3388639

[R6] (Working paper) Wachter, S., Mittelstadt, B., and Russell, C. (2020) Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. http://dx.doi.org/10.2139/ssrn.3547922

4. Details of the impact (indicative maximum 750 words)

Wachter's research on the legal and ethical underpinning of AI systems—particularly her work on a right to explanation, counterfactual explanations, data protection, and algorithmic fairness—has substantially informed policy guidance on AI systems in the UK and Europe, and has also informed significant changes to business practice.

Informing policy guidance on a right to explanation:
The automated decisions enabled by AI systems, while extremely attractive to government and industry because of their time and cost efficiency and accuracy, present a new and significant challenge to regulation, with important implications for equality, fairness and human rights. Wachter's paper **[R1]** on the (lack of a) "right to explanation" of automated decision-making and significant loopholes in the GDPR generated significant public debate, including press coverage, public hearings, and discussion in academic and policy circles, on ways to increase algorithmic accountability. It also informed the House of Commons Science and Technology Committee's public inquiry on 'Algorithms in decision-making' in 2017. Wachter was invited to give evidence to the inquiry on 14 November **[C1a and C1b]**, and an entire section of the Committee's enquiry and report was devoted to exploring the right to explanation, in which Wachter's evidence is cited **[C1c, pp.29–30]**. Wachter also testified as an expert witness before the House of Lords Committee on Artificial Intelligence on October 31, 2017. The House of Lords subsequently submitted amendments to the draft Data Protection Bill, demanding a right to explanation for automated decisions. Four amendments directly speak to Wachter's evidence **[R1]**: amendments 74 (automated processing), 119 (explanation), 134 (automated decision-making), and 183 (the inclusion of a new clause: "Right to information about individual decisions by public bodies based on algorithmic profiling") **[C2]**. These amendments were intended to close loopholes identified **[R1]** and thereby increase algorithmic transparency and accountability.

While the amendments were later dropped, Wachter's research on the right to explanation has put this issue firmly on the policy agenda. It has been cited in over 30 reports on data governance published by influential bodies, including the Royal Society and British Academy, the Nuffield Foundation, the Parliamentary Office of Science and Technology, the European Parliament, and the Council of Europe. Concerning European protections, the Article 29 Working Party, which issues guidance and interprets matters of EU data protection law, fixed the loopholes she raised **[R1]** in their "Guidelines on Automated individual decision-making and Profiling" in 2018 **[C3]**. **[R1]** points out that the GDPR only gives protection for fully automated decisions (i.e. with no "human in the loop")—although this very rarely happens in practice, and "token humans" (i.e. nominal human involvement) can be added to a system in order to avoid such protections. Article 29 fixes this loophole by clarifying that "*To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture*", that is, these tokens *can't* be used as a way to avoid protection under the GDPR. They confirm that these guidelines take account of Wachter's research and list her publications **[R1, R3]** in their recommended reading **[C3, p. 37]**. The guidelines recommend solutions **[C3, pp. 20–22, 24–27]** and offer legal clarifications inspired by Wachter's research to close the extensive legal loopholes on algorithmic accountability she identified.

Changes in business regulation and practice: counterfactual tools to explain AI decisions
Having put the issue AI accountability on the policy agenda, Wachter's subsequent paper **[R3]** – showing that it *is* possible to provide meaningful explanation of so-called black-box decisions by providing "counterfactual explanations" – has indicated a practical way forward for both policy

makers and industry. The guidelines for good explanations issued by the Information Commissioner's Office **[C4]**, the UK's data protection regulator, are heavily influenced by this work; as is the guidance on automated decision-making issued by the Article 29 Working Party citing her work on counterfactuals as good examples of explainable AI **[C3]**.

Following extensive engagement with industry, Wachter's work has also contributed to significant changes to business practice. Companies have come under increasing public and political pressure to increase transparency of automated decision-making, and Wachter's work on counterfactual explanations **[R3]** currently offers the only practical means to do so in a human-understandable way. It has thus been seized on by industry as a solution to the problem, and is becoming increasingly used by developers and businesses around the world.

In their 2019 White Paper on "Issues in AI Governance" **[C5]**, Google cites Wachter's work **[C5, p. 23]** on counterfactuals **[R3]** as a means by which AI system operators can better understand a system's outcomes. Google had already implemented counterfactual explanations in an interpretability interface in September 2018, releasing the "What-If Tool" for use by consumers, businesses, and developers. Designed for Google's TensorFlow, one of the leading machine-learning frameworks, the tool is intended to make "it easier for a broad set of people to examine, evaluate, and compare machine learning models - whether you're a developer, a product manager, a researcher or a student" **[C6]**. Google acknowledges that the tool is directly underpinned by Wachter's research **[R3]** by providing a means for ordinary consumers to understand complex "black-box" decisions **[C6]**.

Sky, for example, has said that "*Understanding how models arrive at their decisions is critical for the use of AI in our industry. […] With tools like What-If Tool […] our data scientists can build models with confidence, and provide human-understandable explanations.*" **[C7].** Google has also implemented counterfactual explanations into their cloud services, making this available to tens of thousands of businesses **[C7].** The Head of Google Cloud AI's Division, has summed up the benefits of implementing counterfactual explanations at Google, saying that: "*it's really important for societal reasons and fairness reasons and safety reasons*" **[C8]** – that is, when it comes to trusting the technology to make important decisions that affect people's lives, in an accountable way.

Other businesses that have implemented counterfactual explanations in their products include IBM, Accenture, Vodafone South Africa (in its "Just 4U" payment plans) [**Corroborator 1]**, and the drone insurance company Flock. Far from being "impossible to understand", black-box decisions about people's finances, medical diagnosis, hiring decisions, university admissions, criminal justice (etc.) can now be made human-understandable **[R2]** as an industry standard, by using counterfactuals as a means of understanding what has / has not been important in coming to a particular automated decision or prediction.

Changing business practice: accountability tools for detection of AI bias

AI systems rely on training data to create their models: we know that (gender, ethnicity etc.) biases in these data can significantly skew models, leading to discrimination against protected groups. While aligning judicial understandings of fairness with AI models has been difficult **[R5]**, Wachter's paper **[R6]** presented a practical means of reconciling these two approaches. In December 2020, an AI accountability toolkit on "fairness detection" was released in SageMaker, Amazon's machine learning service **[C9].** The tool implements 'conditional demographic disparity' (a concept coined and described in **R6**) as a baseline statistical measurement that allows testing of fairness in automated systems by detecting bias in training data. As Amazon explains, "*This metric is useful for exploring the concepts of direct and indirect discrimination and of objective justification in EU and UK non-discrimination law and jurisprudence*" **[R5, R6] [C9].** That is, by aligning with the European Court of Justices' 'gold standard', the tool enables the contextual approach to judicial interpretation of bias practiced under EU non-discrimination law.

Through significant engagement with both policy makers and industry – across the three areas of a right of explanation for automated decisions **[R1, R2]**, counterfactual explanations as a

means to accomplish this **[R3, R4]**, and detection of bias in AI systems **[R5, R6]** – Wachter's work on transparency and accountability of AI systems is now affecting millions of people in all sectors. Important loopholes in European protections on automated decision making have been closed, and industry and consumers now have, for the first time, a means to intuitively understand the reasons for automated decisions, including ways to assess whether and why they have been discriminated against.

5. Sources to corroborate the impact (indicative maximum of 10 references)

C1: House of Commons Science and Technology Committee, 'Algorithms in decision-Making':

- **a.** Wachter's written evidence
- **b.** Wachter's oral evidence 14 November 2017
- **c.** The Committee's report, 15 May 2018 https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/351.pdf

C2: Data Protection Bill [HL], Third Marshalled List of Amendments to be moved in committee of the whole House, 24th October 2017. (page 9 amendment 74, page 21 amendment 119, page 28 amendment 134, page 46 amendment 183). https://publications.parliament.uk/pa/bills/lbill/2017-2019/0066/18066-III(Rev).pdf

C3: Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679', revised and adopted 6th February 2018. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053

C4: Information Commissioner's Office report: 'Explaining Decisions made with AI', https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence/

C5: Google White Paper, 'Perspectives on Issues in AI Governance', 22nd January 2019. https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf

C6: Google cited and implemented Wachter's work on 'Counterfactual Explanations' in their 'What-If' TensorFlow interface, https://pair-code.github.io/what-if-tool/faqs/ and https://www.groundai.com/project/the-what-if-tool-interactive-probing-of-machine-learning-models/1 11/09/2018.

C7: Increasing interpretability of AI with Explainable AI, by Tracy Frey, Director, Product Strategy & Operations, Cloud AI, 21 November 2019. https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-ai-explanations-to-increase-fairness-responsibility-and-trust

C8: 'Google tackles the black box problem with Explainable AI', by Leo Kelion, 24 November 2019: https://www.bbc.co.uk/news/technology-50506431

C9: Wachter's paper cited in Amazon's SageMaker Developer guide [p.587]: https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-data-bias-metric-cddl.html

Corroborator 1: Head of Data, Services and Consumer Regulation at Vodafone Group