Institution: University of Hertfordshire		
Unit of Assessment: 30 – Philosophy		
Title of case study: Improving the Design of Intelligible Artificial Intelligence at Microsoft		
Period when the underpinning research was undertaken: 2015 – 2020		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s): Constantine Sandis	**Role(s) (e.g. job title):** Professor of Philosophy	**Period(s) employed by submitting HEI:** September 2015 – present
Period when the claimed impact occurred: 2017 – 2020		
Is this case study continued from a case study submitted in 2014? N		

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

Through an ongoing collaboration with Microsoft, Sandis's research has contributed to a change in perspective across the company on how to understand 'explainable' Artificial Intelligence (AI) and the moral and legal right to explanation. This has impacted on Microsoft's ethics standards which inform their design of responsible AI; influenced the work of teams working on several products across Microsoft; achieved wider reach beyond Microsoft through non-academic publications, industry and public talks, and through the secondary impact of industry research by others.

2. Underpinning research (indicative maximum 500 words)

Sandis' research conducted at the University of Hertfordshire since 2015 has been instrumental in outlining the relations between understanding, knowledge, intelligibility, and explanation, with a particular focus on the enabling conditions for understanding others (2015a, 2015b, 2016, 2017, 2019, 2020).

Understanding Others

Sandis argues that understanding others does not in any way involve accessing information concerning their 'inner mental states', be it through their verbal reports or, *per impossibile*, by looking inside their minds or brains (2015b, 2016, 2019). This popular conception, which can be traced back to the empiricist philosopher John Locke, is fuelled by the idea that understanding is ultimately a matter of retrieving information that is 'stored' in our minds (2019). On Sandis' alternative conception, whose origins lie with Hegel, Wittgenstein, Collingwood, and Anscombe, the popular 'mindreading' view should be replaced by an account of understanding others that is only achievable through the sharing of behaviour, experiences, and practices (2015a, 2019).

AI Intelligibility

The extension of this view to Artificial Intelligence (AI) has radical consequences for the problem of Explainable AI (XAI), namely that of how to best design AI systems in order to provide users with intelligible explanations of the systems' behaviour. Sandis' philosophy of understanding (see 2a) entails that 'black box' approaches to XAI are misguided in their assumption that we can only explain AI decisions by uncovering the algorithm hidden behind them, and the related goal of building systems that are capable of describing their inner programming in a human-like manner (the equivalent of explaining our actions by describing neural events): if AI could speak, we couldn't understand it (2017).

Sandis argues in his forthcoming book that current industry views focussing on explainable AI are much too narrow and fail to take into account what making an AI system *intelligible* to a "user" actually amounts to (Sandis, C., *Understanding Others: Humans, Animals, & Machines,* forthcoming Yale University Press). In his capacity as Research Consultant to Microsoft, Sandis has accordingly maintained that by focusing on algorithm transparency and natural language programming, the company (and wider industry) approach misses—and at times even interferes with—interactive and counterfactual ways that could improve users' understanding of these complex systems.

XAI Myths

Sandis collaborated with Professor Abigail Sellen, Principal Researcher at Microsoft Research, Cambridge (MRSC) on XAI during his research and consultancy visiting periods at MRSC in 2017 and 2018 (see section 4). Sandis and Sellen (forthcoming) demonstrate that neither algorithm/machine learning transparency nor 'natural language' programming adequately deliver either the (a) XAI or (b) moral transparency required to offer those whose lives are significantly affected by decision-making in increasingly automated domains (ranging from financial to medical technology). In their place they offer a contextualist approach for designing intelligible AI systems so that (a) and (b) above may be achieved through human-computer interactions supplemented with counterfactual analysis (as opposed to the system rendering itself intelligible to a passive 'user'). They build on Sandis' earlier work (esp. 2015b and 2019) to bust three myths of XAI, namely (i) that understanding equals transparency, (ii) that intelligibility is about explanation, and (iii) that the relevant explanations are universal.

The results have implications for designers and developers of AI systems, but also moral and legal discussions of how we should conceive of the 'explanation' of automated decisions. There is currently no legally-binding right to an explanation of automated decisions that affect us (e.g. in terms of health, finance, and justice). Pressure (e.g. from The Turing Institute) for this to change focuses on the moral and proposed legal nature of such a right,but presumes the 'explanation' in question to take the (i-iii) form that Sandis and Sellen reject as myth. XAI is to be achieved not through the release of algorithm and machine learning data but via design and development that is better geared towards interactive and counterfactual intelligibility.

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

3.1 Sandis, C. (2015a). Verbal Reports and 'Real' Reasons: Confabulation and Conflation. *Ethical Theory and Moral Practice,* 18, 267–280. https://doi.org/10.1007/s10677-015-9576-6

3.2 Sandis, C. (2015b) "If Some People Looked Like Elephants and Others Like Cats": Wittgenstein on Understanding Others and Forms of Life. *Nordic Wittgenstein Review*, 4, 131-153. https://doi.org/10.15845/nwr.v4i0.3372

3.3 Sandis, C. (2016) Period and Place: Collingwood and Wittgenstein on Understanding Others. *Collingwood and British Idealism Studies*, 22(1), 171-98. https://www.ingentaconnect.com/content/imp/col/2016/00000022/00000001/art00008

3.4 Sandis, C. (2017) 'If an Artwork Could Speak: Aesthetic Understanding After Wittgenstein', in (ed.) G. Hagberg, *Wittgenstein on Aesthetic Understanding.* Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-319-40910-8

3.5 Sandis, C. (2019), 'Making Ourselves Understood: Wittgenstein and Moral Epistemology', *Wittgenstein-Studien*, Vol. 10, Iss.1, 242-260. https://www.degruyter.com/journal/key/WGST/10/1/html

3.6 Sandis, C. and Sellen, A. (forthcoming), 'Myths of Intelligible AI', draft of paper to be sent to *Communications of the ACM*.

Note: This paper was due to be submitted for publication in 2020 but due to the pressures of the COVID-19 pandemic on work by Sandis's Microsoft collaborator, this has been delayed to 2021. However, the research itself was completed in 2017-20 and underpins much of the work done with Microsoft during that period, so it is included here. A copy is available on request.

Evidence of Quality: 3.1-3.5 peer-reviewed.

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

Sandis's 2015 publications (**3.1**; **3.2**) led to his being invited by Abigail Sellen (Deputy Director and Principal Researcher at Microsoft Research, Cambridge) to present his work at an AI Workshop at Microsoft in May 2016. This paper "sparked many excellent discussions" and was "a really stand-out talk of the day" [**5.1**] As a result, Sandis was invited to join Microsoft's Research Lab in Cambridge as a Research Consultant on AI and Intelligibility for the following

periods (Sellen notes that three months is the maximum time allowed in any one year for such a collaboration):

- July – September 2017: Visiting Researcher with the Human Experience Design (HXD) and Human-Computer Interaction (HCI) teams
- October – December 2018: Research Consultant on AI and Intelligibility with the HCI, HXD and Machine Intelligence and Perception (MIP) teams
- October – December 2019: AI Consultant
- Sandis has also presented twice at the Microsoft Faculty Summit at the company's Headquarters in Redmond, USA, in 2017 and 2019.

Through this ongoing collaboration, Sandis's work has: impacted on Microsoft's ethics standards which inform their design of responsible AI; influenced the work of teams working on several products across Microsoft; achieved wider reach beyond Microsoft through non-academic publications, industry and public talks, and through the secondary impact of industry research by others. Sellen writes that:

> ***"The reach and significance of Sandis' expert advice and its impact on Microsoft is indisputable… [I]t is contributing to a change in the entire company's perspective on how to understand explainable AI and the moral and legal right to explanation"*** [**5.2**].

Impact on ethics standards at Microsoft

Sandis's research has had a direct impact on Microsoft's central AI, Ethics, and Effects in Engineering and Research (AETHER) committee. According to Sellen: "*Senior leadership relies on Aether to make recommendations on responsible AI issues, technologies, processes, and best practices. Its working groups undertake research and development, and provide advice on rising questions, challenges, and opportunities. Ultimately, AETHER's recommendations inform the approach of every single Microsoft employee and the research, design, and engineering of every single Microsoft product*" [**5.2**]. Sellen reports that Sandis's work (both collaborative work with Microsoft, and his previous publications) have been directly informing the committee and, via it, Microsoft's understanding of the core company principle of Transparency (a term which is used as shorthand for the view that AI systems should be understandable) and how to go about ensuring its implementation. She further writes that the research has: "*been fed into the company's ethical standards for intelligible AI, whose rules and guidelines must be adhered to by every single company employee worldwide. As a response to this work, AETHER is evolving to make space for human-centric ICT: not just algorithmic explanations, but design for a more interactive approach. These are large, ongoing, changes in the company's general approach to ethical AI. They have already begun to affect Microsoft's approach to design and engineering. This will impact every single new Microsoft product. The number of worldwide end-users is several millions. Given Microsoft's ambition to lead the tech industry in the realm of Ethical AI, the company's ultimate aim is for these changes to affect the tech community at large. Sandis' conceptual and philosophical points about AI intelligibility inform a very crucial part of this huge and exciting change*" [**5.2**].

Impact on product design

Sandis' initial work with the HXD group in 2017 drew directly upon his research on mutual understanding (2015b) to inform the group's work on human-agent interactions and automated conversation. Through intellectual advice, shared readings, and co-authoring, his expertise helped the group challenge and explore assumptions surrounding human and nonhuman agency, and had, according to the MSRC Senior Researcher for the group at the time "*an instrumental role in ongoing research and development around automated conversation and conversational bots and agents*". He states that as a result, the emphasis of the team's R&D "*shifted… more towards language as the machinery for mutual understanding/ action, rather than a communication channel to be encoded and decoded*" [**5.3**].

In 2018 Sandis contributed to a cross-project machine-learning study run by Richard Banks (Principal Design Manager, MSRC) to share insights into AI intelligibility from across various machine learning projects related to XAI, including large scale and global Microsoft technologies such as **Project Tokyo** (see below); **Project Athens** (launching in 2021 to help game developers and designers understand game agent training); **Healthcare** (launched in 2020 to provide insight into how AI interprets complex signals regarding health); **Project Talia** (launching in 2021 to support care delivered through online psychotherapy programs); **Laser Eyes** (HoloLens launched in 2020 to provide developers with information about what the user is looking at). The study "*drew out insights, learnings, and approaches towards intelligibility… mapped for future use across all projects, and as a way of communicating them with the wider company. All members of the study learned from the work of other members. Accordingly, Sandis' impact on the work of the above…teams and projects is ongoing and diffuse*" [**5.2**].

Sandis took part in the early stages of two Microsoft intelligibility studies relating to Project Tokyo, an emergent cutting-edge technology whose general release aims to deliver AI-enabled technologies to help partially-sighted and blind people to better navigate the world. It augments the user's own capabilities via orientation cues, an external display of the system state, and access to different experience modes. Sandis took part in two intelligibility studies involving blind children, which explored how these 'users' and bystanders understand the system's capabilities. These consisted of (i) a mixed-methods study on how to improve performance via feedback loops (ii) a deep analysis how an agent information source integrates into people's sense-making activities. Both took the approach to interactive intelligibility developed by Sandis and Sellen's collaborative research at Microsoft. Defined via the technological studies themselves, this approach had "*a tangible influence on the outcome of these studies*" [**5.2**]. As Sellen says, "*the potential of these 'other ways of knowing' about a social situation to completely transform the everyday experience of blind and partially-sighted people is life-changing*" [**5.2**]. A pilot study with blind children in UK schools showed that participants "*are acquiring new social skills through the use of this technology*" [**5.2**]. The Project Tokyo system is now being integrated with existing Microsoft technologies for the visually impaired, with internal release in summer 2021, with the expectation being a wider deployment by the end of the year [**5.2**].

Reach beyond Microsoft

Sandis's research has been cited in several industry publications by Microsoft researchers [**5.4**] and he has also jointly authored a number of papers [**5.5**]. A Microsoft Senior Researcher writes: "*In co-authoring, [Sandis] continues to have an impact on both the development work and scholarship in this area*" [**5.3**]. Based on their collaborative research at Microsoft, Sandis and Sellen have written an industry paper [**3.6** above] exposing various myths of intelligible AI and arguing that 'transparency solutions' need to be replaced by design for a more interactive approach to human/machine understanding. Publication of this paper and that of a corresponding non-academic article for the Microsoft Research magazine *Things We've Learnt About*, were planned for 2020, but delayed due to COVID-19.

During his consultation periods, Sandis published his ideas as short pieces on high profile platforms such as *The Conversation* (over 10,000 reads) [**5.6a**], *Medium* (960+ followers) [**5.6b**], and *The Philosophers' Magazine* (readership c.2,000) [**5.6c**] drawing considerable attention to them via social media platforms such as Twitter (9,000+ followers @csandis) and LinkedIn (500+ connections). These short publications enabled Sandis to efficiently communicate his findings to various time-constrained Microsoft researchers, designers, and engineers he met both at Cambridge and around the globe. They also disseminated Sandis' ideas beyond Microsoft, thereby generating additional reach beyond industry, through wider media coverage and invited public talks. As a consequence, Sandis has:

- Been profiled in the international business magazine EXAME, in an article on tech enterprises recruiting philosophers (December 2017) [**5.7a**], and interviewed by award-winning investigative journalist Peter Warren for an episode of his PassW0rd radio

programme on 'Ethics in Technology', London's Resonance 104.4 FM (14 November 2018) [**5.7b**].
- Given an invited keynote talk on his work with Microsoft at a Birkbeck Open Day, to demonstrate the impact that philosophy can have to potential applicants and their parents (February 2018).
- Spoken at a British Academy workshop on AI & Posthumanities, alongside speakers from DeepMind and the Alan Turing Institute [**5.7c**].

Effects of COVID-19 and continuing work

A further visiting period in 2020, and the 2020 Redmond Faculty Summit where Sandis was due to speak on AI Intelligibility, were cancelled due to the effects of COVID-19 on the MSRC lab and Microsoft more generally, but collaboration has continued remotely. Concluding on the impact of Sandis's work overall, his principal Microsoft collaborator writes that:

> ***"The change is a large and therefore gradual one, but will eventually influence everything we do at Microsoft across the entire company, and beyond"*** [**5.2**].

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

5.1 Letter from Professor Abigail Sellen, Deputy Director and Principal Researcher, Microsoft Research Cambridge, 2017.
5.2 Further letter from Abigail Sellen, 2020.
5.3 Letter from Alex Taylor, former Senior Researcher, Microsoft Research Cambridge, 2017.
5.4 Citations in industry-facing articles by Microsoft employees:
- Harper, R., Rintel, S., Watson, R., and O'Hara, K. (2017). The 'interrogative gaze': Making video calling and messaging 'accountable'. Pragmatics 27:3, 319-350. http://doi.org/10.1075/prag.27.3.02har
- Ahmed, S., Balasubramanian, H., Stumpf, S., Morrison, C., Sellen, A. and Grayson, M. (2020). Investigating the Intelligibility of a Computer Vision System for Blind Users. In: IUI '20: Proceedings of the 25th International Conference on Intelligent User Interfaces. (pp. 419-429). New York: ACM. http://dx.doi.org/10.1145/3377325.3377508
- Thieme, A., Cutrell, E., Morrison, C., Taylor, A., and Sellen, A. (2020). Interpretability as a dynamic of human-AI interaction. Interactions 27:5, 40-45. https://doi.org/gb34

*Note: Sellen writes that Sandis's work "had a tangible influence on the outcomes of these studies" [**5.2**] referring to Ahmed et. al (2020) and Cutrell et. al. (2020).*

5.5 Joint publications with Microsoft employees.
- Sandis, C., & Harper, R. (2018). Wittgenstein and Communication Technology: A conversation between Richard Harper and Constantine Sandis. Philosophical Investigations, 41(2), 241-262. https://doi.org/10.1111/phin.12188
- The AI of the Beholder – non-academic article on Medium. https://constantinesandis.medium.com/the-ai-of-the-beholder-303ecac5fc9d

5.6 Selected examples of reach to non-academic audiences:
a) 'We don't want AI that can understand us – we'd only end up arguing' in *The Conversation*, 21 August 2017. https://theconversation.com/we-dont-want-ai-that-can-understand-us-wed-only-end-up-arguing-82338
b) https://constantinesandis.medium.com/
c) Twelve columns, e.g. 'Philosophy, In a Sense: Enchanted Action' on real and artificial agency. https://archive.philosophersmag.com/philosophy-in-a-sense-enchanted-action/
For full list see: https://archive.philosophersmag.com/?s=sandis

5.7 Selected examples of media coverage and public talks:
a) EXAME December 2017, pp.73-4
b) http://www.futureintelligence.co.uk/fis-password-radio-show/
c) https://posthumanities.co.uk/a-i/