Smart Technologies and Our Sense of Self: Going Beyond Epistemic Counter-Profiling

Sylvie Delacroix[1, 3] and Michael Veale[1, 2, 3]

[1]Birmingham Law School, University of Birmingham
[2]Dept of Science, Technology, Engineering and Public Policy, University College London
[3]The Alan Turing Institute, London

Both authors contributed equally to this work

The final version of this draft will be available in *Law and Life in the Era of Data-Driven Agency*, O'Hara & Hildebrandt (eds.), f'coming 2019, Edward Elgar Publishing Ltd.

Preprint date: **April 12, 2019**

This chapter focuses on the extent to which sophisticated profiling techniques may end up undermining, rather than enhancing, our capacity for ethical agency. This capacity demands both opacity respect—preserving a gap between the self we present and the self we conceal—and an ability to call into question practices that are ethically wanting. Pushed to its limit, the smooth optimisation of our environment may prevent us from experiencing many of the tensions that otherwise prompt us to reconsider accepted practices. An optimally personalised world may not ever call for any 'action' as Hannah Arendt describes it.

Can systems be designed to personalise responsibly? Greater time and research needs to be invested in designing a range of viable 'perspective widening' tools, as many such tools either burden users with little guarantee of meaningful engagement, or underestimate the extent to which individuals' preferences are themselves malleable. Any approach that tries to predict what users might like, or what might change their views, risks the same pitfalls as any other form of personalisation. Instead, we argue that the most promising avenue is to push for diverse uses of newly developed systems, and measure those systems' success at least partly on that basis. Inviting appropriation and repurposing would help keep users engaged in systems of data collection and profiling. This will not be a straightforward task: sometimes it will be in tension with traditional measures of success and performance. Yet the increasing integration of algorithmic systems in society requires us to widen our understanding of agency beyond a narrow, decontextualised focus on passive consumption preferences.

Contents

1 Introduction

Pervasive and complex digital profiling practices have moved from niche concern to global debate.[1] International scandal has raged around the fine-grained and supposedly influential political profiling undertaken by firms such as Cambridge Analytica, powered by covertly obtained data and facilitated by the lax respect for privacy and data protection by technology giants. The chair of the relevant UK House of Commons oversight group, the Digital, Culture, Media and Sport (DCMS) Committee, claimed the online broadcast of one of their evidence sessions was 'the biggest ever live-streamed audience that parliament has ever had' (Helm 2018). While social media content is now under increased scrutiny, similar and perhaps more invasive industries are emerging to process data collected from sensors in homes, environments, and even on our bodies. Given these developments, such scandals are likely to be relatively commonplace in the years to come.

To approach these problems, we must define them. Researchers and practitioners in this field have tended to work with some canonical problem framings to think about societal implications. A typical framing sees a large data controller[2] using machine learning techniques on personal data to build a model to predict something consequential: whether someone will pay back a loan, buy a product, or commit a crime. They query this model with new input data relating to a particular individual to make or support a decision concerning them. The opacity of the decision system limits an individual's ability to contest it (for example, under

[1] The GDPR, in Article 4(4), defines profiling as 'any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements'.

[2] A data controller is a natural or legal person who, jointly or alone, 'determines the purposes and means of the processing of personal data' (Article 4(7), GDPR).

equality law). That individual might have an instrumental or an intrinsic claim to a right to transparency, as recognised by freedom of information laws; algorithmic information provisions in administrative law, such as recent French law (Edwards and Veale 2018); the common law 'duty to give reasons' in English law (Elliott 2011; Oswald 2018); and transparency rights and obligations in relation to automated decision-making systems present in European data protection law (Edwards and Veale 2017).

These rights have occupied much of the debate around algorithmic systems. They broadly attempt to rectify epistemic imbalances: a powerful entity knows or understands something concerning an individual that they themselves do not. This knowledge asymmetry can be exploited in a way that the individual is powerless to prevent. All the rights and obligations mentioned above (from transparency to freedom of information) stem from a general effort to mitigate the effects of unprecedented epistemic imbalances. These efforts are far from new in computing: explanation facilities have a long history in expert systems research (Guidotti et al. 2018). However such facilities have primarily been aimed at decision-support users, rather than affected individuals (Binns et al. 2018). The perceived need for 'algorithmic accountability'—has motivated a range of new research communities to consider the extent to which such 'downstream transparency' can—and should (Weller 2017)—be provided. Alongside the latter have been cautious voices, arguing that the transparency ideal is rife with limitations (Ananny and Crawford 2016), and that solely focusing on transparency rights may further individualise responsibility for ensuring systems are socially aligned, a 'transparency fallacy' which fails to give individuals meaningful control just as the 'notice and choice' paradigm failed in the past (Edwards and Veale 2017).

Information flows between actors are key to Mireille Hildebrandt's understanding of the risks of computational profiling in *Smart Technologies and the End(s) of Law* (Hildebrandt 2015, hereafter The End(s)). Hildebrandt draws on a range of scholarship, including G. H. Mead and Helmuth Plessner, to consider how an individual's behaviour -and process of identity formation- is shaped by a range of mutual expectations. Such expectations are multilateral in nature: individuals do not just internalise a set of rules that constitute a role, nor seek to get inside another actor's head, but must intuitively internalise expectations that an entire social framework generates of themselves in relation to those around them, and of those around of them in relation to each other. Non-human agents must build similar understandings if they are to integrate seamlessly in our lives too, and generally must do so in a different way than we can. We must do the same to navigate a world that contains sophisticated non-human actors capable of analysis and adaptation. Hildebrandt is concerned that, faced with highly granular profiling systems (insofar as they achieve their intended predictive functions at all), individuals will be unable to effectively anticipate the way they will be read, and therefore are left in an unequal and problematic situation which may harm them as well as the broader social fabric.

To address this, in *The End(s)* Hildebrandt proposes the mechanism and frame of 'counter-profiling'. This consists of 'conducting data mining operations on the behaviours of those

that are in the business of profiling, whether 'those' are humans, computing systems or hybrid configurations' (Hildebrandt 2015: 223) to help redress the imbalances generated by machine learning systems. Hildebrandt does not see this as a necessarily individual activity, but one which could be undertaken collectively, akin to the means by which media organisations currently hold important systems to account in constitutional democracies. Such counter-profiling organisations have been proposed at different scales, from the use of 'super-complaint'–style powers in Article 80 of the GDPR to empower NGOs (Edwards and Veale 2018), platforms to facilitate stories on machine learning systems in journalism (Trielli et al. 2018), or the collective use of data protection rights with the purpose of aggregating the results and potentially 'reverse engineering' consequential systems (e.g. Mahieu et al. 2018; Palmetshofer and Semsrott 2018).

Counter-profiling might generate an array of types of knowledge. In recent years, there has been a heightened focus on procedures to obtain information about the decision-logics of predictive systems due to the tendency of such systems to utilise undesirable proxies for use in decision-making. Typically, a decision-system might be using only seemingly non-sensitive data as input, but inferring some latent class or variable which has not been explicitly measured, such as race, disability, or social status, and using this to, in part, determine outcomes (Barocas and Selbst 2016; Calders and Žliobaitė 2013). At an individual level explanation facilities are another way in which counter-profiling can be operationalised. Often, these proceed by creating a simplified version of a model (such as the area around a decision-point), or snapshot part of its logic (such as a sensitivity analysis) to provide further information to a user. This information might serve a variety of purposes: from increasing trust (or a sense of due process), to enabling users to react and take certain actions, or to detect 'bugs' and errors in a deployed system.

Such forms of counter-profiling seem focused on *epistemic* imbalances. One side 'knows' more than the other, even if this knowledge is not expressed in a semantic or otherwise interpretable form (Burrell 2016): rectifying this imbalance is the core purpose of such counter-profiling endeavours. The latter are mostly concerned with the models themselves, rather than also turning to consider the sociotechnical systems they inhabit and co-create.

Yet when we profile each other in the world, the heuristics and intuitions we develop along the way go beyond what is required to build a simple model of a single person. The problem is that explanation facilities within machine learning systems proceed from precisely this 'simple model' perspective, aiming to simplify a system in a way that allows humans to understand it. Even where these explanation facilities are designed to prescribe reaction or foster response (eg for an individual to know how she should change her input data, if she can, to get different results from a machine), their overarching objective is still to simplify a single software object so that a human can understand or interpret it.

In *The End(s)*, by contrast, Hildebrandt describes counter-profiling as a wider endeavour. Individuals are not assumed to be simply seeking information about biases, or explanations of software: instead they might be working together (as they do in purely human interactions) to

understand 'hybrid configurations' (p. 223). The focus is on re-enabling mutual anticipation, rather than simply balancing knowledge. As noted by Hildebrandt, efforts to regain mutual anticipation across society that are stuck in the simplistic "[opening] the skull of another person' (p.222) perspective seem likely to stall before they make significant progress.

We too are doubtful that a focus on epistemic rebalancing will meaningfully address the societal issues stemming from sophisticated profiling systems. In this chapter, we elaborate on one challenge that epistemic counter-profiling fails to consider: to what extent is our commitment to equality imperilled by our insatiable appetite for optimisation via cheap, ubiquitous sensors and actuators?

2 Pervasive profiling as social cruelty

To understand the fundamental nature of the commitment to equality we are concerned about, it is helpful to start by considering instances when this equality is most clearly negated. What do rape, genocide, torture and slavery have in common? They are, following Andrea Sangiovanni, all paradigmatic instances of *social cruelty*.

> '[S]ocial cruelty involves the unauthorised, harmful, and wrongful use of another's vulnerability to attack or obliterate their capacity to develop and maintain an integral sense of self.' (Sangiovanni 2017)

How can the seemingly benign, profile-based endeavour to optimise the way we spend our time (along with the food we eat, the people we meet, and the news we read) be even *remotely* connected to practices as abhorrent as genocide? It helps to start by considering what makes genocide distinctively wrong. Sangiovanni (2017) articulates this by reference to:

> 'the reasons for mass murder, which are grounded in an ideology that singles out a group of people as deserving extermination in virtue of who they are [...] Those affected can no longer appear in public without fear or maintain a social self that is (partially) defined or controlled by them.'

No matter what kind of person the member of a group targeted for genocide may have been aspiring to become, being targeted in such a way renders those aspirations and efforts of self-definition irrelevant (just like the efforts of the person subjected to rape or torture).

Even in mundane circumstances, retaining some sense of ownership over the way one projects oneself, both socially and through one's body, is never easy. To do so requires, minimally, a to-and-fro movement between the process of definition of one's 'self' from *without*, such as the effects of natural events and human encounters, and from *within*—how one appropriates such events and encounters. This to-and-fro movement is easily imperilled. Events such as a grave illness (or prosecution) can leave one with the sense that one no longer knows how to continue (and hence appropriate such events), given who one was (Delacroix 2018).

In those cases, the threat to the to-and-fro movement is exogenous. What would a threat from *within* look like?

We posit that the invisible, profile-based optimisation of our environment may, in the longer term, undermine our capacity to develop and maintain an integral sense of self *to such an extent* as to fall within the scope of Sangiovanni's definition of social cruelty. This endangers our commitment to *moral*—not merely epistemic—equality.[3] How so? Just like the slave-owner (or the authorities behind a genocide) not only does not need, but precludes any active input from the person the slave (or member of a group targeted for genocide) is trying to be, data-controllers can 'read' me, build a profile that accurately anticipates my upcoming moves, desires and risks without any need for active input on my part. In the data-controller's 'eyes' -and consequently in the eyes of other users of the profile-based technology, I am merely becoming the person anticipated by that profile

Given the impact of this definition of self *from without*, my ability to resist or contest this extraneous definition, already under pressure, will be increasingly compromised as those profile-based, ubiquitous computing applications leverage recent advances in affective computing (rebranded in recent years as 'Emotional AI'). Profiling systems which anticipate desires and preferences based on physiological indicators of nascent emotions may all too easily compromise the 'counter-process' of definition *from within* by making any effort of appropriation seemingly redundant. Rather than experiencing the tension(s) between nascent feelings (such as those leading to anger) and one's meta-preferences or aspirations (for instance the type of parent one strives to be), the individual may be presented with an instantly 'optimised' environment that either removes or masks the cause of one's upcoming anger. While such optimisation would probably make for smoother interpersonal relationships, at least in the short term, it may also leave us with poorer versions of ourselves, ones that are deprived of the chance to learn from and grow through the experience of such tensions.[4] Most worryingly, this fine-tuned optimisation of our environment makes it considerably more likely that we will end up conforming to the profile-based, extraneous definition of ourselves, thus turning such profiles into self-fulfilling prophecies.

Hildebrandt does allude to this concern, stating that such smart technologies 'may thus 'read' our emotions before we have a chance to develop our own reflection and response to our own emotional state [...] [i]f we cannot contest the way we are being "read" and steered and thus if we cannot resist the manipulation of our unconscious emotional states, we may lose the sense of self that is pre-conditional for human autonomy. Not having a chance to develop feelings, we may actually become the machines that smart technologies take us to be' (Hildebrandt 2015: 71–72). Here, we wish to unpack this insight further, and propose mechanisms and practices which might help avoid this undesirable end.

[3] As for why the latter is best understood as a commitment to the absence of social cruelty, see Sangiovanni (2017).

[4] On the value of anger, see Srinivasan (2018).

3 Affective computing and the new behaviourism

Affective computing, more recently rebranded as 'emotional AI', is one of the technologies that play a central role in the optimisation endeavours above. It is defined as 'computing that relates to, arises from, or deliberately influences emotions' (Picard 1997: 249). One of the domain's early proponents, Rosalind Picard, argued strongly for the consideration of the functions that emotions play in processing and problem-solving—that computers 'do not need affective abilities for the fanciful goal of becoming humanoids; they need them for a meeker and more practical goal: to function with intelligence and sensitivity' (Picard 1997: 247). As well as considering the role of emotions *in* computing, a major strand of affective computing surrounds affect *detection* through a range of modalities including face, posture, blood volume pulse, skin conductivity, vocal tone, spoken or written text, and more (see generally Calvo et al. 2015). Fuelled in particular by heavy commercial interest in using these systems to understand consumers in a marketing context, research in this field has been plentiful and well-funded. Such research has gone far beyond the question of whether digital affect detection is even possible (e.g. Healey and Picard 1998) to attempt to computationally parse 'micro-expressions': involuntary and near-invisible muscle movements thought to betray us; expressions researchers believe are 'capable of revealing the actual emotions of subjects even though such leaks are unintentional' (Xu et al. 2017). These are involuntary leakages or flags which might indicate deception: laughing at an unfunny joke, smiling while angry, or the like. Such flags have been of research interest for nearly half a century (Ekman and Friesen 1969), and, with the ease of computation and cheapening of high resolution, high frame-rate cameras that can capture these transient involuntary signals, has led to a range of datasets and predictive benchmarks (see e.g. Yan et al. 2014). Richer and more confident understanding of affect might also be gleaned from fusing multiple sources of information such as posture (Gunes and Piccardi 2005) or data from wearables (Gunes and Hung 2016).

Coupled with what Antoinette Rouvroy terms 'data behaviourism'—the belief (rightly or wrongly) that human actions can be well-anticipated through inductive methods applied to large amounts of data (Rouvroy 2013)—this would seem to leave data subjects quite open to pre-emption and interpretation by data controllers in ways data subjects have limited practical ability to control or mitigate. Such pre-emptive approaches are already receiving publicity and drawing concern. In 2018, a spate of news stories surrounded a new patent filed by the firm Amazon, which owns and controls the 'Echo' or 'Alexa' voice assistant ecosystem. While many firms patent technologies that never go on to enter production, they are often deemed indicative of a firm's thinking or internal research spend. This patent, *Voice-based Determination of Physical and Emotional Characteristics of Users*, sought to understand whether users were experiencing an 'emotional abnormality' and aimed to take some form of corrective action. The illustrative example provided (Figure 1) shows a woman who appears to be ill or upset being asked whether she would like a recipe for chicken soup.

While the validity of this approach requires strong behaviourist assumptions, in practice

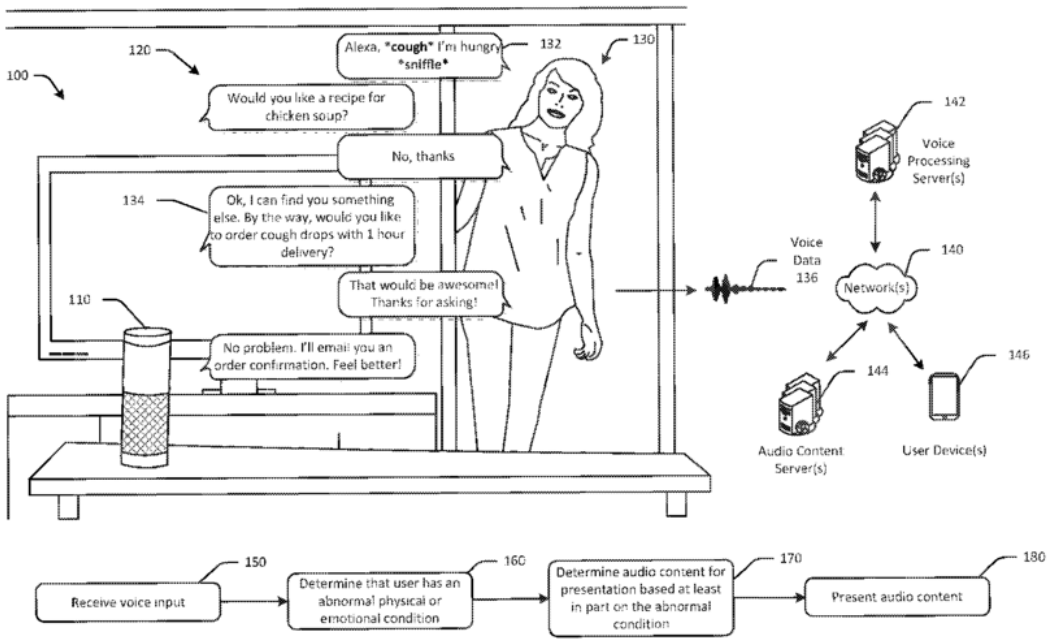

Figure 1: Patent no US010096319, Amazon Technologies Inc. Source: United States Patent and Trademark Office, www.uspto.gov (public domain).

incorrect or miscalibrated inferences will have tangible effects on how individuals act and are expected to act—and indeed on their very capacity for 'action'.

As that through which we 'make our appearance' (Arendt 1958)—revealing who—rather than what—we are, action is inherently unpredictable. Hannah Arendt reminds us of this in at least two respects. As the vehicle through which each and every citizen exercises her agency—her capacity to start something new—true action presupposes an element of surprise. 'The fact that man is capable of action means that the unexpected can be expected from him, that he is able to perform what is infinitely improbable' (Arendt 1958: 178). Action is also unpredictable to the extent that we cannot anticipate exactly what kind of self we will reveal through it, nor how that self will be read and narrated by those who surround us.

This unpredictability inevitably generates anxiety, and Arendt deplores the progressive shrinking of the public sphere (*polis*) (concomitant with the widespread withdrawal to the private pursuit of economic interests) as Modernity's regrettable coping mechanism. The advent of profile-based optimisation tools that aim to mitigate (or possibly cancel altogether) the unexpectedness inherent in action goes one significant step further, as an 'unpredictability coping mechanism'. By potentially short-circuiting *thought* itself (that space where *I* am called to balance conflicting feelings, aims and aspirations), the smooth, profile-based optimisation of our environment may ultimately render us incapable of *judgment*. Without the latter, without some ability to question deeply internalised habits of thought and action, there

is no *polis*: only a large number of what Arendt refers to as the pre-political 'homo faber'—who are focused on the production of artefacts—working side by side at best (or, at worst, a series of thoughtless 'Eichmanns') (Arendt 1994).

4 Fostering self-definition

The downstream regulation of profiling practices, with its ex-post remedies, such as transparency rights after-the-fact, seems an unlikely answer to the dystopian future described above. Not only might these remedies be too little, too late, but these rights after-the-fact also require a capacity for reflexive action. Reflection theorists have highlighted the need to create appropriate settings and 'scaffold' simple information provision in order to enable it to have a transformative effect on individuals, and it is unclear whether such characteristics are built into the rights and provisions discussed above (Slovák et al. 2017). Furthermore, it is possible that highly granular profiling processes themselves might systematically undermine the ability to reflect. It is furthermore highly unclear what these ex-post remedies would be ex-post *of*: an individual profiling instance? Systematic exposure and behaviour change over many months, or years? A sudden, regretful action that an individual attributes to her environment? There may well be individual uses for these remedies, but they are no panacea, and at best they will need to be augmented with other means and measures.

Here we instead focus on upstream design interventions. We see two main avenues through which political and ethical agency might be preserved alongside some of the benefits that personalisation and pre-emption might bring. We review these in the context of some connected efforts in the relevant research literature—particularly in human–computer interaction—and highlight their opportunities, as well as their unresolved gaps and pitfalls, which are in dire need of future research and practice.

4.1 Passive interventions: Profiles in perspective

Much interface design has centred on ideas of 'seamlessness' (Weiser 1994), where the functioning of a system (such as utilised profiling practices) are hidden to the user to better let them focus on the task they want to achieve. While this can help users with certain tasks, in many cases, it can also obscure the way that a given system works, and individuals might lose perspective as a result.

To counterbalance this, some researchers have explored means of giving users greater perspective on how their experience of a system relates to others' experience of a system. In some forms, this might resemble the counter-profiling efforts we were perhaps unduly critical of above. Give users a better understanding of how a system works, and they might want to work differently with it, against it, or customise it to their needs (Ekstrand and Willemsen 2016). Yet it is not required for an individual to understand how a system works and 'perceives' them (epistemic equality) to be given the opportunity to reflect, and something to

reflect upon. Here, building on *The End(s)* (see p. 222–3), we consider interventions which do not seek to *explain* the profiling system, but instead provide perspective in other ways and with more consideration of the sociotechnical context and other actors involved. Might such perspective provide the moral jolt required to keep the to-and-fro of self-definition going?

One interesting approach has been to focus on many different users at once, and giving individuals perspective upon how *others* might be seeing a platform or recommender system. Webster and Vassileva (2007) propose a news recommender system with a twist: it is accompanied by a visual map of yourself in relation to other users in your social network, whom you can place closer or farther away from yourself to change their own influence on what you see, as well as take a glimpse into the kinds of recommendations they will be getting. Kang et al (2016) propose a similar system for Twitter users to explore content 'popular just beyond a user's typical information horizon'. Here, the focus is less on 'why did I get this recommendation?', than on 'what are others using this system experiencing?'.

This approach appears promising, but also contains pitfalls. While some primarily public platforms like Twitter already support basic ways to see the timelines of other users, it is not easy or intuitive for users to access this functionality. Moreover, much content is private in nature, so putting yourself in someone else's shoes is limited by the streams and viewpoints you are authorised to access, for example on platforms like Facebook. Furthermore, when moving beyond the recommender system paradigm, systems can be personalised in more functional ways. The way that individuals have configured their smart home, or series of smart devices, might be fundamentally incompatible with the profiles of other individuals. Short of engaging in a lifestyle swap, with hardware and software included, it is hard to see how this approach translates neatly into many newly datafied areas of life. Lastly, it is not at all a given that seeing other users—who may indeed be subject to personalisation just as deeply, and their own reflective capacities equally challenged—will provide the stimulus needed to help a user retain ownership over the process of identity formation and self-definition.

A different approach surrounds users not with the data or information of others, but accumulated information on *themselves* which they may not have mentally logged or been aware of. The *Balancer* web extension, for example, displays a stick figure in the browser carrying a load (with a carrying pole) which would become more lopsided the less ideologically diverse the media consumed was (Munson et al. 2013). The use of this tool appeared to have a small observed effect on encouraging more diverse media consumption. Other tools aim to encourage users to directly take perspectives, such as develop pro/con lists (Kriplean et al. 2012a), reflect upon other users' perceptions (Kriplean et al. 2012b), or mapping out user comments onto axes of values to prompt users to navigate across them more broadly (Faridani et al. 2010). What users might *not* have done could also be of interest, and in some ways, more challenging to convey. Tintarev et al (2018) presented users with visualisations of their profiles, and found evidence that some were effective in helping users identify 'blind-spots'.

In the physical world, this approach echoes *personal informatics*, also referred to as self-tracking or the 'quantified self' movement. This area primarily concerns fitness wearables at

present but does have a range of applications (and adherents) more broadly. Research in this field has emphasised that users do not reflect on the data trails they collect as a result of using these devices and systems in straightforward and uniform ways, but instead their practices are highly task and goal dependent (Li et al. 2011). While relevant data with reflective potential might be able to be logged, to encourage heterogeneous users to reflect upon it in ways that are usable, useful and meaningful to them is likely to be challenging. While there may be ways to introduce information into a user's environment (e.g. through ambient interfaces displaying data designed to be 'processed in the background of awareness' (Wisneski et al. 1998)), whether such passive approaches, often based on lights or sounds and lacking depth or complexity (e.g. Houben et al. 2016) would be effective at promoting moral reflection is a question for empirical research.

4.2 Active interventions: Surprises and appropriation

If passively illustrating highly personalised systems to users is not enough to encourage reflection and behaviour change, might more active interventions hold promise? Are there ways of building reflectivity–inducing tools that would prompt end-users to question the deeply ingrained habits of thought and action that make their behaviour as end-users suitable for profiling in the first place?

In the research field of recommender systems, there has long been interest in understanding how or why simple metrics, such as accuracy on training data, do not appear to relate fully with users' satisfaction with systems. This suggests that other values or constraints, aside from accuracy, may be pertinent in the design of such systems. Two are of particular interest in this context. The first is the notion of *serendipity*. Coined by Horace Walpole in 1754, over time serendipity has become used to describe an accidental, unplanned, but valuable discovery, often in a scientific context (Merton and Barber 2011). In human-computer interaction, it has generally been seen as a novel prediction that triggers a positive emotional response from a user. Another, related notion is *unexpectedness*. Unexpectedness does not restrict itself to *novel* (totally unseen) predictions or actions, but instead draws upon a model of what the user is expecting. Those predictions that are outside a user's expected range, but not so far outside to be irrelevant, are considered 'unexpected' (Adamopoulos and Tuzhilin 2014).

The notions of serendipity and unexpectedness as they have been explored in the recommender systems literature share two problematic assumptions in relation to the issues described in this paper.

First, they take an individual's preferences as simply not fully known, rather than malleable or not fully-formed. A similar assumption is shared by microeconomics, where even complex consumer preferences are static and do not depend on the options at hand (cf. Dietrich and List 2013). This might be reasonable in trivial domains, but quickly becomes controversial in consequential cases, particularly morally charged ones. There is a wide array of literature in the field of political science on how individuals' value-laden preferences change in response

to actions, messages, sources of information, and under different conditions of knowledge (for a review, see Druckman and Lupia 2000). It is these types of techniques which might be deployed, even automatically, to give individuals moments of moral reflection, or to attempt to stir them from the routines they take for-granted. Those designing and maintaining predictive systems might think themselves archaeologists, uncovering individuals' preferences, but might they be sometimes better described as architects shaping and developing them (Bettman et al. 1998; Gregory et al. 1993)?

Understanding how individuals' preferences change would be an ethically challenging area of research and practice. Intervention-based, habit forming research is relatively uncontroversial when the aim is known by the participant and agreed to—for example, to form habits commonly thought of as healthy, such as changing eating patterns or exercising more. Where issues of concern have heavier moral or political dimensions, it becomes less so. Facebook attracted significant criticism for its experimental intervention, without informed consent, to test the 'emotional contagion' hypothesis, whereby individuals' emotional states are influenced by those around them. Researchers attempted (and claimed to succeed in) manipulating individual behaviour by altering the composition of the content they were exposed to (Kramer et al. 2014). These kind of methods—the A/B testing common to technology firms today—form the basis of 'agile' development of software systems. Such practices have been highlighted in the context of how they transform privacy concerns and necessary governance (Gürses and van Hoboken 2018), or how they might engage concerns typically found in research ethics (Bird et al. 2016). Quasi-experimental methods (which require no active random intervention) might be useful, but given that they tend to require particular fortuitous setups, they are notoriously difficult to apply generally, and would therefore be unlikely to see easy uptake with regards to the day-to-day development of systems.

Secondly, both serendipity and unexpectedness are typically introduced as constraints aimed at maximising user satisfaction or perceived utility. Serendipity in particular (given its potential emotional underpinnings) has yet to be considered as a possible reflexivity-inducing, habit-distancing tool meant to foster agency, not just user satisfaction. To understand how serendipity and political/ethical agency may be related, it is useful to consider the extent to which habit all too easily compromises the latter. Our capacity for what Arendt calls 'judgment', i.e. our capacity to call into question widely accepted practices when they are wanting, indeed presupposes an ability to step back from the habitual, to query seemingly routine values or practices. This capacity cannot be taken for granted: challenging as it is to maintain such critical distance in an 'offline world', it becomes even more so when the technology we interact with is designed to be habit-inducing to a much greater extent than the 'natural' objects that structure our quotidian habits.

To what extent can we—and should we—design the profiling systems we so extensively rely on in our daily lives to sometimes 'work backwards', and periodically prompt us to reconsider the very traits and habits that feed our respective profiles? Can we design such systems in such a way as to shake us out of deeply ingrained habits through surprise? Such suggestions

have been made in particular for decision-support systems (Delacroix 2018), and it is worth considering how this logic might extend into ubiquitous computing applications. For such surprises to have that agency-enabling, habit-shaking effect, they would have to go beyond 'trivial' prediction failures, and lead us to reconsider our understanding of the world, and our place within it. Given this likely axiological component, would such 'surprise-based' interventions amount to covert paternalistic interventions? Not if their goal is not so much to 'nudge' us towards a particular choice (Sunstein and Thaler 2003), but instead to enable us to creatively make choices for ourselves, rather than blindly adhering to old habits. Instead of narrowing down choice to a particular goal or target, the aim could be to open it up more widely, and celebrate diverse uses of technology. Here one may usefully leverage recent research on the factors that impact upon individuals' different levels of creativity (Zabelina et al. 2011)—these factors include 'fluency' and 'flexibility'—and endeavour to counter the effects of routinisation with applications that are not that dissimilar to those used in the context of art and mathematics (Boden 2010) as well as business (Adam et al. 2006).

Yet for such efforts to be genuinely agency-enabling, one would need to move beyond the overwhelmingly passive role which end-users are typically endowed with, and design systems that carve them an active role within the learning loop (this approach is sometimes referred to as 'interactive machine learning' or 'IML' (Wallach and Allen 2008). An explicit requirement to keep monitoring the results of the learning process, combined with a demand for regular, creative input on the part of end-users, indeed has the potential to not only improve the system's learning performance; it may also keep thoughtless torpor at bay by encouraging an 'ethical feedback loop'.

Related to this are a range of calls that ask designers not to design products for imagined users, but to give individuals space to define and purpose technologies themselves (see eg Hildebrandt 2019). Dix (2007) calls this 'designing for appropriation', arguing designers should take a principled approach to allow for interpretations, provide visibility and avoid overemphasis on seamlessness, expose the intentions of particular features or routines, seek to support users rather than control them, allow coupling, chaining and modular expansion, and encourages sharing and learning. This echoes calls in a range of fields, such as mobile health (mHealth) to consider technologies which go beyond paternalistic prevention to encourage 'self experiments and reflective practice' (Churchill and Schraefel 2015). This is accompanied in regulatory terms by the nascent but growing "right to repair" movement, which seeks to enshrine the ability to tinker, fix and have access to replacement parts for objects that are increasingly difficult for users to open and mend (Koebler 2017). All of these directions appear to head for active efforts for users to engage in the systems they are part of, and the introduction, growth or enforcement of any one of them may bring spillover effects to support the causes we have outlined above.

5 Conclusion

Our appetite for control and predictability is far from new. The sophisticated profiling endeavours attempting to personalise our choices can be read as merely extending the scope of this long-standing appetite. Personalisation does have clear benefits: aside from reducing cognitive load, it is essential to developing technologies which are more tailored and accessible to individual needs, and navigating and making sense of the swelling mass of digitised information. Yet the extent to which, just like any tool (and perhaps more so than any other tool), it will end up *changing* its users must be considered critically. Here, we have argued that our current regulatory focus on addressing epistemic imbalances (e.g. through 'explainable' systems) is, on its own, inadequate, and additional approaches and tools are needed.

This chapter focused on the extent to which sophisticated profiling techniques may insidiously imperil the fragile movement of to-and-fro that is essential to any person's endeavour to continuously re-define her sense of self in a way that commands respect, both from herself and from others. This respect for each and every person's need to preserve a gap between the self that we present to the world and the self that we conceal is at the heart of our commitment to moral equality. It presupposes a certain degree of opacity, an opacity threatened by our supposedly strengthening ability to infer future desires and preferences from past behaviour and/or physiological indicators. But the challenges raised by profiling techniques go further than the need to preserve opacity respect. It is our very ability to step back and call into question existing practices (what Arendt would refer to as 'judgment') that is potentially at stake. Pushed to its limit, the smooth optimisation of our environment may indeed prevent us from ever experiencing the tensions or emotional discomforts that prompt us to step back and reconsider practices we took for granted. An optimally personalised world may not ever call for any 'action' as Arendt understands it.

Can systems be designed to personalise responsibly? Greater time and research needs to be invested in designing viable 'perspective widening' tools that leverage the knowledge we have acquired on the mechanisms underlying habit-formation and transformation. Encouraging users to reflect on the use of systems by themselves and others—for example, by making usage logs compellingly visible—has been suggested, but heavily burdens the individual with little guarantee of meaningful engagement. More active approaches, such as the concepts of serendipity and unexpectedness developed in recommender systems and information retrieval seem initially promising, but also come with their own sets of flaws. Primarily, they are still modelling and catering to user satisfaction, and underestimate the extent to which individuals' preferences are themselves malleable. Any approach that tries to predict what users might like, or even what might change their views, risks the same pitfalls as any other form of personalisation we have described.

Instead, we argue that perhaps the most promising avenue seems to be to push technologists to aim for their systems to be diversely used, and measure success at least partly on that basis. Inviting appropriation and repurposing, and exploring how it can be supported,

would help keep users engaged in systems of data collection and profiling. This would not be a straightforward task, and may at times be in tension with traditional measures, such as accuracy. Yet simple, technical fixes that sit neatly alongside existing evaluation paradigms are unlikely to be fixes at all. For designers and developers to appreciate that is the first important step to understanding how to navigate this new risk in a world of 'smart' technologies.

6 Acknowledgements

Michael Veale acknowledges funding from the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/ M507970/1]. Thanks go to Mireille Hildebrandt and a further anonymous reviewer for comments on this work which shaped the final version.

References

Adam, F., P. Brezillon, S. Carlsson and P. Humphreys (2006), *Creativity and Innovation in Decision Making and Decision Support*, London: Ludic Publishing.

Adamopoulos, P. and A. Tuzhilin (2014), 'On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected', *ACM Transactions on Intelligent Systems and Technology*, **5** (4), 1–32.

Ananny, M. and K. Crawford (2016), 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability', *New Media & Society*, accessed at https://doi.org/10.1177/1461444816676645.

Arendt, H. (1958), *The Human Condition*, Chicago: The University of Chicago Press.

Arendt, H. (1994), *Eichmann in Jerusalem: A Report on the Banality of Evil*, New York: Penguin.

Barocas, S. and A. D. Selbst (2016), 'Big Data's Disparate Impact', *California Law Review*, **104**, 671.

Bettman, J. R., M. F. Luce and J. W. Payne (1998), 'Constructive Consumer Choice Processes', *The Journal of Consumer Research*, **25** (3), 187–217.

Binns, R., M. V. Kleek, M. Veale, U. Lyngs, J. Zhao and N. Shadbolt (2018), "It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions', *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'18)*, accessed at https://doi.org/10.1145/3173574.3173951.

Bird, S., S. Barocas, K. Crawford, F. Diaz and H. Wallach (2016), 'Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI', *Presented at the 3rd Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT/ML 2016), 18 November 2016, New York City, New York*, accessed at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2846909.

Boden, M. (2010), *Creativity and Art: Three Roads to Surprise*, Oxford: Oxford University Press.

Burrell, J. (2016), 'How the machine 'thinks': Understanding opacity in machine learning algorithms', *Big Data & Society*, **3** (1), accessed at https://doi.org/10.1177/2053951715622512.

Calders, T. and I. Žliobaitė (2013), 'Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures', in B. Custers, T. Calders, B. Schermer, and T. Zarsky (eds), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 43–57.

Calvo, R. A., S. D'Mello, J. Gratch and A. Kappas (2015), *The Oxford Handbook of Affective Computing*, Oxford University Press.

Churchill, E. F. and M. c. Schraefel (2015), 'mHealth + Proactive Well-being = Wellth Creation', *Interactions*, **22** (1), 60–3.

Delacroix, S. (2018), *A Vulnerability-Based Account of Professional Responsibility*, accessed at https://doi.org/10.2139/ssrn.2840864.

Delacroix, S. (2018) 'Taking Turing by Surprise? Designing Digital Computers for morally-loaded contexts' *arXiv preprint arXiv:1803.04548 [cs.CY]*, accessed at https://arxiv.org/abs/1803.04548.

Dietrich, F. and C. List (2013), 'Where do preferences come from?', *International Journal of Game Theory*, **42** (3), 613–37.

Dix, A. (2007), 'Designing for Appropriation', in *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 2*, Swindon, UK: BCS Learning & Development Ltd., pp. 27–30.

Druckman, J. N. and A. Lupia (2000), 'Preference Formation', *Annual Review of Political Science*, **3** (1), 1–24.

Edwards, L. and M. Veale (2017), 'Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not The Remedy You Are Looking For', *Duke Law and Technology Review*, **16** (1), 18–84.

Edwards, L. and M. Veale (2018), 'Enslaving the algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?', *IEEE Security & Privacy*, **16** (3), 46–54, accessed at https://doi.org/10.1109/MSP.2018.2701152.

Ekman, P. and W. V. Friesen (1969), 'Nonverbal Leakage and Clues to Deception', *Psychiatry*, **32** (1), 88–106.

Ekstrand, M. D. and M. C. Willemsen (2016), *Behaviorism Is Not Enough: Better Recommendations through Listening to Users*, ACM Press, pp. 221–4.

Elliott, M. (2011), 'Has the Common Law Duty to Give Reasons Come of Age Yet?', *Public Law*, (Jan), 56–74.

Faridani, S., E. Bitton, K. Ryokai and K. Goldberg (2010), 'Opinion space: a scalable tool for browsing online comments', in *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, New York, New York, USA: ACM Press, p. 1175.

Gregory, R., S. Lichtenstein and P. Slovic (1993), 'Valuing environmental resources: A constructive approach', *Journal of Risk and Uncertainty*, 7 (2), 177–97.

Guidotti, R., A. Monreale, F. Turini, D. Pedreschi and F. Giannotti (2018), 'A Survey Of Methods For Explaining Black Box Models', *arXiv Preprint 1802. 01933v [cs. CY]*.

Gunes, H. and H. Hung (2016), 'Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block', *Image and Vision Computing*, **55**, 6–8.

Gunes, H. and M. Piccardi (2005), 'Affect Recognition from Face and Body: Early Fusion vs. Late Fusion', in *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, IEEE, pp. 3437–43.

Gürses, S. and J. van Hoboken (2018), 'Privacy after the Agile Turn', in E. Selinger, J. Polonetsky, and O. Tene (eds), *The Cambridge Handbook of Consumer Privacy*, 1st ed., Cambridge University Press, pp. 579–601.

Healey, J. and R. Picard (1998), 'Digital processing of affective signals', in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98*, IEEE, pp. 3749–52.

Helm, T. (2018), '"Was the Brexit poll compromised? We may need a public debate about that"', *The Guardian*, 14 April, accessed 15 April 2018 at http://www.theguardian.com/uk-news/2018/apr/14/damian-collins-mp-interview-need-reform-electoral-law-digital-age.

Hildebrandt, M. (2015), *Smart Technologies and the End(s) of Law*, Cheltenham, UK: Edward Elgar.

Hildebrandt, M. (2019), 'Privacy As Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning', *Theoretical Inquiries in Law*, **19** (1), accessed at https://doi.org/10.2139/ssrn.3081776.

Houben, S., C. Golsteijn, S. Gallacher, R. Johnson, S. Bakker, N. Marquardt, L. Capra and Y. Rogers (2016), 'Physikit: Data Engagement Through Physical Ambient Visualizations in the Home', in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, pp. 1608–19.

Kang, B., N. Tintarev, T. Höllerer and J. O'Donovan (2016), 'What am I not Seeing? An Interactive Approach to Social Content Discovery in Microblogs', in *Lecture Notes in Computer Science*, pp. 279–94.

Koebler, J. (2017), 'Source: Apple Will Fight "Right to Repair" Legislation', accessed 16 May 2018 at https://motherboard.vice.com/en_us/article/mgxayp/source-apple-will-fight-right-to-repair-legislation.

Kramer, A. D. I., J. E. Guillory and J. T. Hancock (2014), 'Experimental evidence of massive-scale emotional contagion through social networks', *Proceedings of the National Academy of Sciences of the United States of America*, **111** (24), 8788–90.

Kriplean, T., J. Morgan, D. Freelon, A. Borning and L. Bennett (2012a), 'Supporting reflective public thought with considerit', in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work - CSCW '12*, New York, New York, USA: ACM Press, p. 265.

Kriplean, T., M. Toomim, J. Morgan, A. Borning and A. Ko (2012b), 'Is this what you meant?: promoting listening on the web with reflect', in *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, New York, New York, USA: ACM Press, p. 1559.

Li, I., A. K. Dey and J. Forlizzi (2011), 'Understanding my data, myself: supporting self-reflection with ubicomp technologies', in *Proceedings of the 13th International Conference on Ubiquitous Computing - UbiComp '11*, New York, New York, USA: ACM Press, p. 405.

Mahieu, R. L. P., H. Asghari and M. van Eeten (2018), 'Collectively Exercising the Right of Access: Individual Effort, Societal Effect', *Internet Policy Review*, **7** (3), accessed at https://doi.org/10.14763/2018.3.927.

Merton, R. K. and E. Barber (2011), *The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science*, Princeton University Press.

Munson, S. A., S. Y. Lee and P. Resnick (2013), 'Encouraging Reading of Diverse Political Viewpoints with a Browser Widget', in *ICWSM*, aaai.org, accessed at http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6119/6381.

Oswald, M. (2018), 'Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **376** (2128), 20170359.

Palmetshofer, W. and A. Semsrott (2018), 'Get involved: We crack the Schufa!', accessed 15 April 2018 at https://okfn.de/blog/2018/02/openschufa-english/.

Picard, R. W. (1997), *Affective Computing*, Cambridge, MA: The MIT Press.

Rouvroy, A. (2013), 'The end(s) of critique: Data behaviourism versus due process', in *Privacy, Due Process and the Computational Turn*, London: Routledge, pp. 157–82.

Sangiovanni, A. (2017), *Humanity without Dignity: Moral Equality, Respect and Human Rights*, Cambridge, MA: Harvard University Press.

Slovák, P., C. Frauenberger and G. Fitzpatrick (2017), *Reflective Practicum: A Framework of Sensitising Concepts to Design for Transformative Reflection*, ACM Press, pp. 2696–707.

Srinivasan, A. (2018), 'The aptness of anger', *The Journal of Political Philosophy*, **26** (2), 123–44.

Sunstein, C. R. and R. H. Thaler (2003), 'Libertarian Paternalism Is Not an Oxymoron', *The University of Chicago Law Review. University of Chicago. Law School*, **70** (4), 1159.

Tintarev, N., S. Rostami and B. Smyth (2018), 'Knowing the unknown: visualising consumption blindspots in recommender systems', in *Proceedings of the 33rd ACM/SIGAPP Symposium On Applied Computing. Pau, France April 9-13, 2018*.

Trielli, D., J. A. Stark and N. Diakopolous (2018), *Algorithm Tips: A Resource for Algorithmic Accountability in Government*, accessed 15 April 2018 at https://perma.cc/3ER3-49FE.

Wallach, W. and C. Allen (2008), *Moral Machines: Teaching Robots Right from Wrong*, Oxford: Oxford University Press.

Webster, A. and J. Vassileva (2007), 'The keepup recommender system', in *Proceedings of the 2007 ACM Conference on Recommender Systems - RecSys '07*, New York, New York, USA: ACM Press, p. 173.

Weiser, M. (1994), 'The World is Not a Desktop', *Interactions*, **1** (1), 7–8.

Weller, A. (2017), 'Challenges for transparency', in *2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*.

Wisneski, C., H. Ishii, A. Dahley, M. Gorbet, S. Brave, B. Ullmer and P. Yarin (1998), 'Ambient displays: Turning architectural space into an interface between people and digital information', in *International Workshop on Cooperative Buildings*, Springer, pp. 22–32.

Xu, F., J. Zhang and J. Z. Wang (2017), 'Microexpression Identification and Categorization Using a Facial Dynamics Map', *IEEE Transactions on Affective Computing*, **8** (2), 254–67.

Yan, W.-J., X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen and X. Fu (2014), 'CASME II: an improved spontaneous micro-expression database and the baseline evaluation', *PloS One*, **9** (1), e86041.

Zabelina, D., D. L. Robinson, D. Council, Michael, J. R and K. Bresin (2011), *Patterning and Non-patterning in Creative Cognition: Insights From Performance in a Random Number Generation Task*, accessed at https://doi.org/10.1037/a0025452.