Ethical Governance for Collective Adaptive Systems.

Mark Hartswood and Marina Jirotka

1. Introduction

This article explores how information technologies in crystallising social change stir up important human values related concerns and how value-sensitive approaches can help foresee and forestall the more serious risks to important values such as personal autonomy, privacy, and social inclusion, as well as being a source of creative engagement for design.

To begin we explore why human values have moved towards centre stage in digital technology design and unpick the multiple relationships that exist between technology and values. Secondly we examine a series of specific examples of how technology interacts with values or valued practices and show their potential relevance to Hybrid, Diversity-Aware Collective Adaptive Systems (HDA-CAS)[2]. Thirdly, we consider how value-sensitive design, as part of a Responsible Research and Innovation (RRI) agenda, can help articulate stakeholder values as an explicit part of the design process. Finally, we consider moving beyond the state-of-the-art in these areas towards a framework for the ethical governance of HDA-CAS.

2. Technology and human values

As digital technologies have become deeply intertwined with all aspects of our lives, including work, leisure and our friendship and familial relations, then thinking through the subtle ways they interact with human values has emerged as a significant strand of human factors research (Sellen et al 2009; Friedman et al, 2006; Knobel and Bowker, 2011). This increased emphasis on values has a number of roots. One relates to a growing focus on 'user-experience' as a key element of technology design and the recognition that accommodating cultural values improves a technology's acceptance (e.g. Marcus, 2000). Another has to do with how search, mobile and social technologies change how knowledge, transparency and accountability are socially distributed, and in doing so alter power dynamics across a wide range of relationships – interpersonal, doctor-patient, citizen-state, consumer-corporation (Kobsa, 2009; Mort et al, 2003; Lanier, 2013). Sellen et al (2009) draw these perspectives into a broader picture locating the need for a values-aware approach within five contemporary techno-cultural trends:

1. **End of interface stability.** No longer a single, well defined means of interacting with computers which are now more ubiquitous and embedded.
2. **The growth of techno-dependency.** Deep economic and cultural dependency on complex technical infrastructures and services with complex interactions, failure points and vulnerability.
3. **The growth of hyper-connectivity.** Attention consuming (always on, always connected) yet generative and powerful.
4. **The end of the ephemeral.** Traces of our activities are no longer fleeting, discarded, forgotten, but are recoverable, creating a digital footprint that has value but also problematics.
5. **The growth of creative engagement.** Computers as tools for expression and creativity rather than just mechanical problem solvers. We can all produce content, create programmes (e.g. IF This Then That), express opinions, create analysis, sift

[2] Hybridity involves the participation of human individuals in the system; diversity-awareness highlights that the operation of the CAS should be sensitive to the identification of behaviour of the individuals and collectives in the system.

information for our friends, changing the balance between traditional patterns and modes of cultural production and consumption.

Sellen et al's (2009) analysis might be critiqued for clinging to traditional HCI concerns as their emphasis lies close to the interaction between individuals and computers. To balance this one might add a sixth category - especially relevant to Collective Adaptive Systems - the emergence of social computing – which is also associated with a complex moral terrain in ways that we unpick in later sections.

It is useful at this point to clarify what we mean by 'human values', and to draw upon sociological literature to guide how we interpret their bearing on technology, and vice versa. Friedman et al in advocating 'value sensitive design' (which we discuss more fully later) as an approach to incorporating a values perspective into design, use a working definition of 'social values' as referring to the importance attached to things by a group or an individual (Friedman et al, 2006). This reading is close to our common sense interpretation that somebody (or some group) with an espoused set of values will tend to take a particular view, be involved in certain practices or behave in a particular way under given circumstances. In this sense, values equate to something that is close to the idea of a personal or shared ethos. Scholars exploring the relationship between values, culture, technology and social behaviour warn against developing these intuitive readings of values into positions of determinism. The two forms of determinism that are at stake here are 'technical determinism' – the idea that technologies straightforwardly shape values, and the converse position, 'cultural determinism', where social values are seen as central to technology selection (and the shaping of social patterns more generally) (Ackermann, 1981; Swidler, 1986).

Rejecting deterministic accounts partially rests on the recognition that social values do not have a singular or primary role in social processes. For Ackermann social values have multiple aspects: they are evaluative and invoked in "situations of choice", they are enacted through practice and cultural expression, and they play a role in sustaining patterns of social relations. Similarly for Swidler values are seen as resources for organising action, rather than determining social behaviour (Swidler, 1986). By implication, the intersection of technology with values is not singular and unambiguous, but multiple and complex. Thus the technology itself might be the thing that is valued, or symbolic of wider values, or implicated in the disruption (or creation) of valued practices. Several examples these types of relationships are given below.

3. The politics of automated mechanisms and embedded algorithms.

Internet search is one of the commonest, most palpable encounters we have with a CAS-like social machine where sophisticated analytics shape search listings based upon continuously updated profiles of the web and of user activity. Search has become ethically and politically charged because of the powerful role it plays in ordering our experience of the virtual world and its physical world referents. Cultural assumptions are frequently and covertly embedded within the mechanisms for ranking and ordering search results in ways that are hard for users to discern and untangle (Introna and Nissembaum, forthcoming). This can have the effect of privileging some cultural perspectives while diluting others, for example, when searching in Google on the word 'Cameroon' it is not until the fourth or fifth page of results that a Cameroonian voice can be found (Knobel and Bowker, 2011). Moreover, each search made subtly contributes to everybody's search experience through incremental adjustments to ranking algorithms, sometimes creating an imprint or reflection that can be read as a dominant cultural preference. These can reflect negatively regarded views including racial biases or prejudices (Sweeney, 2013). Prior work examining the ethical principles important for developing and deploying algorithms embedded in computer models may prove a useful starting point for unpicking these types of issue. This work draws attention to the relationships of responsibility between algorithm developers and their ultimate users (Johnson

and Mulvey, 1995), and the significance of, and means of, achieving visibility of mechanisms and embedded values (Fleischmann and Wallace, 2009).

Implications for HDA-CAS: Because algorithms are complex and obscure, mediating collective adaptation can generate suspicion that values have been covertly embedded, or else enable adaptations that are partial or that reflect darker social values. Understandings here could help inform requirements for trust and transparency and thus, be of interest to the provenance work, for example, where there is loss of traceability with aggregation.

4. Values attached to social practices

Values can interfere with the uptake of technologies. For example, the Senegalese were motivated by their government to adopt ("modern") gas burners to replace charcoal stoves in a move to preserve forests and exploit surplus gas. While many units were purchased, few remained in routine use. One reason is that it was hard to use them to make good tea, an important practice in Senegalese family life, although many of the burners were kept for *display* to symbolise progressive values (Ackermann, 1983). A simple reading of this story is that dominant social practices associated with strongly held social values shaped the adoption and use of technology. More subtly, one can see the way that the proffered gas burners figured at different times in relation to different sets of values, sometimes as a bearer of those values, sometimes as a disruption to valued practices, and along an unfolding trajectory hard to discern from the outset.

Implications for HDA-CAS:
These perspectives could be important for work on Incentives and Decision-Making in exploring how HDA-CAS intersect with existing human values for their effective operation. For example, how might interference with values attached to dominant social practices impede participation within HAD-CAS? Or conversely, how do we motivate participation by identifying, building upon and extending existing social practices and their attendant values? Finally, how might we understand the complex interactions and multiple roles technologies play in an evolving value system?

5. Values attached to labour

Technological developments impact in a myriad of ways upon the character and availability of paid labour. Possible effects include, reduction in overall demand for labour, shifts in the availability of different forms of labour, changes to how labour is valued and remunerated, changes in patterns of work, the decline and extinction of some forms of work, and the emergence of new forms. Part of this turns on the sorts of markets enabled by networked technologies. On the one hand new market opportunities are heralded by opening up the 'long tail' of demand (Brynjolfsson & Smith, 2006), but on the other are the dystopian narratives of the erosion of skilled labour and middle class occupations and the concentration of power and wealth with so-called 'siren servers' (Lanier, 2013).

Emerging labour markets driven by crowdsourcing approaches such as the MechanicalTurk, as well as creating new labour opportunities, have also attracted criticism for exploitative practices, (including low wages, and poor reward prospect for skilled contributions), placing the burden of risk with the worker and allowing enticements to participate in fraudulent activities (such as forging reviews and spamming) (Silberman et al, 2010). Although much crowdsourcing involves financial gain (for participants or organisers), others are based on non-monetary forms of value exchange such as a community participation and public acknowledgement and appear less likely to attract negative ethical appraisal (Karpinsky et al, n.d.).

Implications for HDA-CAS:

Important questions for HDA-CAS relate to the motivation and remuneration for crowds – how will the supply of this type of labour evolve along with increasing demand; what will future expectations be on the types of exchanges that are considered fair in digital participation? More subtly, what will be the effects of HDA-CAS on traditional work roles and what sorts of transitions should be anticipated and managed? The distribution of work between the people and the machine has interesting consequences in terms of who/what is responsible if things go wrong.

6. Interpersonal values

How do technologies intersect with values attendant on interpersonal relationships? For example, it is may be useful to use geo-locating services to track friends to enable spontaneous meeting and just-in-time arrangements. But those self-same technologies become technologies of accountability by creating a sense of entitlement to know where a partner is, or why they were offline (Kobsa, 2009). New services then emerge that enable people to lie about their location. (Knobel and Bowker, 2011). A second example is how the etiquettes and moral responsibilities of friendship are transformed by technologies such as Facebook via practices of 'friending' and 'unfriending' etc (Holmes, 2011).

Implications for HDA-CAS
Sensitivities where HDA-CAS touch upon collaborating networks of families and friends – e.g. when attempting to leverage a patient's social network. Understanding in these areas will help inform tradeoffs between transparency and privacy.

7. Social sorting and social exclusion

Social sorting concerns how surveillance technologies enable population to be stratified in ways that shape entitlement or access to services or resources (Lyons, 2003). An example would be corporations minimising their risks by adjusting services according to data they can gather about customers' health statuses. Ambient healthcare technologies are seen to have a high potential for generating detrimental social sorting (Kosta et al, 2010). Social exclusion is a related form of social partitioning whereby peoples' material or social disadvantage prevents them from participating in valued social practices, which can be exacerbated by lack of access to socially enabling technologies (e.g. Valentine et al, 2009).

Implications for HDA-CAS
This may help us understand the risks that **peer profiles** pose as a resource for (unwanted and detrimental) social sorting. A more general point concerns what we understand about the populations who participate in a HDA-CAS (and the populations that are disengaged) and how different sorts of inclusiveness make some statistical measures less reliable.

8. Value Sensitive Design (VSD) and Responsible Research and Innovation

Value sensitive design is an approach to digital technology design that aims incorporates an exploration of stakeholders' values into the design process by making visible values related concerns such as those outlined above. VSD can be seen as belonging to a swathe of approaches contributing to the 'ethical governance' of emerging technology that come together under the banner of Responsible Research and Innovation (RRI). RRI aims to mitigate the negative consequences of technological progress through a mix of ethical practices, regulation and governance appropriate to a scientific domain. These include: foresight, democratisation, public engagement, and the fostering of outcomes-sensitive innovation and design practices (Stahl, 2013, Von Schomberg, 2013 ; Owen et al, 2012).

Value centred design is an approach to technology design that incorporates an appreciation of human values to enable more socially acceptable forms of technology to emerge. It does this in a number of ways: By helping design to proceed in full awareness of what is at stake and

for whom; so that when decisions are taken they are not done so blindly or in ignorance. It also aims for sensitivity to emergent values, and how technologies transform existing value regimes by bringing newly valued practices, objects and capabilities into play. Lastly, but not least, understanding values has a creative potential to inform the sorts of services that would be meaningful to stakeholders, and to help foster desired forms of participation and social behaviour.

VSD employs three forms of investigation to uncover relevant values (Friedman et al, 2002):

1. **Conceptual investigations:** To identify who is affected, in what way, what values are at stake and how trade-offs between competing values might be managed. This is typically a desk-based research that could involve exploring the philosophical basis for concepts such as 'consent' or analysing a technology scenario from the perspective of normative values and existing governance and legislative frameworks.

2. **Empirical Investigations:** Deepens the conceptual investigation by exploring value issues at play in relevant personal, organisational or societal contexts. Empirical investigations might reveal the interplay of values within a setting, or across settings, or it might focus on the sorts of values that are apparent through the use of a particular technology, for example, what are the determinants of privacy preferences in Facebook?

3. **Technical Investigations:** Understand the interplay between technological affordances and specific value regimes. A good example here is the concern raised by Google Glass, which discretely embeds a video camera in a wearable item thereby greatly increasing the potential for unappreciated digital surveillance.

VSD makes an important distinction between 'direct stakeholders', who are recognised within existing design approaches, and 'indirect stakeholders' who are neither immediate users nor developers, nor owners, but whose lives may nonetheless be affected. Drawing again on the Google Glass example, an indirect stakeholder might be someone who is inadvertently caught on video by a Google Glass user. VSD also makes a series of distinctions between designer values, explicitly supported (or designed in) values and stakeholder values, each of which might be distinct and non-overlapping. The aim is to ensure that the explicitly supported values are not merely an embedding of the designer's own values or preconceptions.

9. Beyond the state of the art - towards ethical governance for HDA-CAS

Several questions emerge as how we move beyond the state of the art in the ethical governance of technologies to inform the development of HDA-CAS. Moreover, the specific issue of how an HDA-CAS might be governed raises particular questions of its own: Who will set the rules for HDA-CAS? What will be its constitution? How will power be vested?

The functional design of social machines and the design of their governance mechanisms are intimately linked because social regulation is also key for enabling the sought after social computation (Hendler et al, 2008; Ericson and Kellogg, 2000). It follows that design for emergent social behaviour depends as much upon artfully embedding appropriate social values as it does on the creation of structures and interactions around which the social behaviours cohere. Our contention is that the design of effective social machines needs to take account of the ways values are communicated though structure, and the role they play to underpin social interaction. By way of analogy, modern architectural practice recognises how social values embedded in, or symbolically communicated by, the built environment are important to regulating human behaviour in the physical world (Shah & Kesan, 2007).

Thus, ethical governance for Social Machines has inward and outward facing aspects. Inwardly, the way that the Social Machine is constituted needs to reflect the ethos of the computations (e.g. meeting standards of authorship in Wikipedia), to offer reciprocity for participation and discourage malicious forms of participation. Outwardly, the role that Social Machines play within a wider societal context is important, e.g. preventing social machines from being put to malicious purposes such as their use by spammers to circumventing Captchas, or in promulgating exploitative labour practices (Silberman et al, 2010). This implies that we have a layered and interacting series of governance concerns and with important ethical considerations emerging at each 'level'.

To design governance mechanisms for HDA-CAS we can draw upon well-known governance mechanisms in other domains, including: professional codes of conduct, democratic processes and other structured forms of decision-making and accountability practices. The governance mechanisms of Wikipedia have been extensively studied as a celebrated example of the potential of socially orchestrated knowledge production. An important finding here is how the governance of Wikipedia has evolved in line with the changing demands of its growing scope and sophistication (Aaltonen and Francesco, 2011). One can draw upon the governance literature more generally to survey candidate governance mechanisms, for example, 'polycentric governance' concerns how users of a commons (with appropriate resources for enforcement and transparency) can better co-regulate its exploitation than could a centralised authority (Ostrom, 2010). 'Adaptive governance' approaches developed for socio-ecological systems, are suited to systems whose dynamically unfolding character drives the need for regulatory regimes to co-develop in a responsive way (Hatfield-Dodds et al, 2007).

Thus, ethically sensitive approaches to developing and operating HDA-CAS would draw upon existing approaches, but in all likelihood weave them into new configurations. For example, VSD provides a sound starting point, but the practices of *design* for social machines well may be different to those of more conventional technologies. For social machines design is more like fine-tuning the conditions for emergence, rather than the creation of a rigid working solution based upon a static specification (McBride, 2011). Also, the adaptive and evolutionary properties envisaged for HDA-CAS give us a temporally extended view of *values* and *stakeholders*. So an initial analysis is unlikely to reveal downstream impacts as the HDA-CAS evolves via learning and in response to changes to its external environment.

A values-sensitive approach would need to be responsive to the emerging implications as it adapts and evolves. Lastly, but not least, value-sensitive approaches acknowledge that populations will hold diverse values, and lead us to consider mechanisms to articulate conflicting values and mediate tradeoffs.

References

Ackerman, W (1981) Cultural Values and Social Choice of Technologies, International Social Science Journal, 33 (3) 447–465.

A. Aaltonen and L. G. Francesco (2011) Governing Social Production in the Internet : The Case of Wikipedia ' in The proceedings of 19th European Conference on Information Systems (ECIS).

Brynjolfsson, E., Y. J. Hu, M. D. Smith (2006). From Niches to Riches:

Anatomy of the Long Tail, MIT Sloan Management Review, 47(4), 67-71.

Erickson, T. and Kellogg, W., A. (2000) Social translucence: an approach to designing systems that support social processes. ACM Transactions on Computer-Human Interaction. 7(1) 59-83.

Friedman B., Kahn P., Borning A. (2002). Value Sensitive Design: Theory and Methods. UW CSE Technical Report 02-12-01,

http://www.urbansim.org/pub/Research/ResearchPapers/vsd-theory-methods-tr.pdf.

Hatfield-Dodds, S., Nelson, R., and Cook, D. (2007) Adaptive governance: An introduction, and implications for public policy. Paper presented at the 51st Annual conference of the Australian Agricultural and Resource Economics Society, Queenstown NZ, 13-16 February 2007

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T. and Weitzner (2008) Web science: An Interdisciplinary Approach to Understanding the Web. Communications of the ACM 58(7) 60-69.

Holmes, M. (2011) Emotional Reflexivity in Contemporary Friendships: Understanding It Using Elias and Facebook Etiquette, Sociological Research Online, 16(1) 11 http://www.socresonline.org.uk/16/1/11.html https://mywebspace.wisc.edu/stenerson/web/documents/EthicsPaper.pdf

Introna, L. D. and Nissembaum H. (forthcoming) Shaping the web: Why the politics of search engines matters. The Information Society.

*www.nyu.edu/projects/nissenbaum/papers/**searchengines**.pdf*

Fleischmann, K and Wallace, W., 2009, Ensuring transparency in computational modeling. Communications of the ACM, 52(3) 131-134. http://dl.acm.org/citation.cfm?id=1467278

Friedman, B., Kahn Jr, P. H., and Borning A. (2006) Value Sensitive Design and Information Systems Forthcoming in P. Zhang & D. Galletta (Eds.), *Human-Computer Interaction in Management Information Systems: Foundations. M.E. Sharpe, Inc: NY.*

Johnson, D. G. and Mulvey, J. M., 1995, Accountability and computer decision systems, Communications of the ACM 38(12) 58-64.

Karpinsky, N., Lall, C., Moore, D., and Stenerson M. Ethics of Crowdsourcing (web manuscript)

Kennedy, H. (2012) Perspectives on Sentiment Analysis. Journal of Broadcasting & Electronic Media 56(4) 435-450.

Knobel, C. and Bowker, C. G. (2011) Values in Design, Communications of the ACM, 54(7) 26-28.

Kobsa, A. (2009) The Circles of Lattitude. Adoption and Usage of Location Tracking in Online Social Networking. IEEE International Conference on Computational Science and Engineering, Vancouver, Canada, 2009, 1027-1030.

Kosta, E., Pitkänen O., Niemelä, M. and Kaasinen, E (2010) Mobile-Centric Ambient Intelligence in Health- and Homecare—Anticipating Ethical and Legal Challenges. Science and Engineering Ethics June 2010, Volume 16, Issue 2, pp 303-323

Lanier J (2013) Who Owns The Future? Pub. Allen Lane.

Lyon, D. (ed). (2003) Surveillance as social sorting. Privacy, risk and digital discrimination. Routledge.

N. McBride (2011) From Social Machine to Social Commodity: Redefining the concept of social machine as a precursor to creating new web development approaches ACM Web Science Conference, Koblenz, June 14-17 2011. url: http://www.websci11.org/fileadmin/websci/Posters/155_paper.pdf

Marcus, A. and Gould, E., W. (2000) Crosscurrents: cultural dimensions and global Web user-interface design. ACM Interactions Magazine 7(4) 32-46.

Mort M, May C R and Williams T (2003) Remote doctors and absent patients: Acting at a distance in telemedicine. Science, Technology and Human Values 28(2) 274-295.

Ostrom, E. (2010) Beyond Markets and States: Polycentric Governance of Complex Economic Systems. The American Economic Review, 100(3) 641-672.

Owens, R., Macnaghten, P. and Stilgoe, J. (2012) Responsible Research and Innovation: From science in society to science for society, with society. Science and Public Policy 39, 751-760.

Sellen, A., Rogers, Y., Harper, R. and Rodden T. (2009) Reflecting Human Values in a Digital Age. Communications of the ACM (52)3 58-66.

Shah, R. C., & Kesan, J. P. (2007) *How Architecture Regulates*. Journal of Architectural and Planning Research, 24(4), 350-359.

Stahl, B. C., Eden, G., & Jirotka, M. (2013). Responsible Research and Innovation in Information and Communication Technology - Identifying and engaging with the ethical implications of ICTs. In R. Owen, M. Heintz, & J. Bessant (Eds.), Responsible Innovation (pp. 199-218). Wiley.

Silberman, M. S., Irani, L. and Ross, J. (2010) Ethics and tactics of professional crowdwork. XRDS: Crossroads, The ACM Magazine for Students. 17(2), 39-43.

Swidler, Ann. 1986. "Culture in Action: Symbols and Strategies". *American Sociological Review* 51(2):273–286.)

Sweeney, L. (2013) Discrimination in Online Ad Delivery. Online manuscript. (? Peer reviewed). http://arxiv.org/ftp/arxiv/papers/1301/1301.6822.pdf

Valentine, G., Holloway, S., and Bingham, N. (2002) The Digital Generation?: Children, ICT and the Everyday Nature of Social Exclusion. 34(2) 298-315.

Von Schomberg, Rene (2013). "A vision of responsible innovation". In: R. Owen, M.
Heintz and J Bessant (eds.) Responsible Innovation. London: John Wiley, forthcoming