Accountable Intelligence

Professor Patrick G.T. Healey,
Queen Mary University of London

Professor Alan Bundy
University of Edinburgh

Abstract: A defining characteristic of human intelligence is the ability to account for our actions. When asked, we are normally expected to provide credible explanations of what we did and why. Interestingly, this applies even when processes are involved -such as low-level vision- that we have only limited verbal access to. However, these accounts are not just rational reconstructions, they are routinely tested against the available behavioural evidence and through friendly, and sometimes unfriendly, cross-examination. This paper proposes that fully useful, intelligent machines necessarily require similar abilities i.e., need to be able to produce, defend and negotiate credible accounts of their own actions. This is essential for the adoption of intelligent machines in the home, hospitals, courts and other contexts where the rationale for an action has a critical social, ethical and legal status. Especially where those machines employ algorithms that render any direct interpretation of their processing problematic. This raises some key challenges for the future development of machine intelligence. It requires giving machines the ability to represent and reason about their own behaviour. It requires an ability to communicate the results of that reasoning in a manner commensurate with natural human interaction. It also requires an ability to flexibly revise its accounts where it becomes clear through time or through interaction with its users that its own account of its actions is unreliable or unconvincing; the mark of intelligence is intelligibility.

Human Accountability:

Accountability is a defining characteristic of human intelligence. We expect people to be able to produce accounts of their actions that are both intelligible and recognisable to others (Garfinkel, 1984). These accounts are typically complex interactive performances that build on shared social and physical context. Moreover, these representations of behaviour feed into human accountability in a second sense; the determination of responsibility for particular actions or 'holding people to account'. In everyday interactions this can involve working to make sense of apparently odd or 'unaccountable' behaviour (Robinson, 2016). In institutional contexts it involves using both people's immediate and recorded accounts as documents of what 'actually' happened and often has a specific legal significance (Garfinkel, 1984). In many such cases an account is not only expected but *required*, for legal or professional reasons.

Paradoxically, we expect people to be accountable despite the acceptance that they do not necessarily have declarative access to all of the mechanisms that govern their behaviour. Low-level sensorimotor processes and automated processes such as syntactic processing are not directly available for verbal report. The Cognitive Sciences conventionally distinguish between different forms of knowledge on the basis of what people are able to represent verbally e.g. 'declarative' and 'procedural' (e.g. Anderson, 1976) or 'explicit' and 'implicit' knowledge (e.g. Broadbent and Fitzgerald, 1986).

Although we do not always have direct access to the some of the computational processes that contribute to our actions we are none-the-less able to produce predictively adequate

‘intentional’ models of our own (and others) behaviour (c.f. Dennett, 1987). Directly attending to the problem of representing our own mental processes during complex decision making can improve our predictions of performance (Osman, 2010). Moreover, the act of articulating these accounts of our own and other’s behaviour provides further tests of the intelligibility and reliability of our accounts. Intelligibility is collaboratively tested through specific, structured conversational processes such as clarification and repair. The predictive adequacy of these representations can then be jointly assessed against observed behaviour The demands of sustained interaction ensure that people’s accounts are repeatedly tested for both their intelligibility and reliability over time.

Machine Accountability:

Intelligent machines are not currently accountable in the senses outlined above. Although most systems are designed produce some form of feedback to users they are unable to flexibly adapt their accounts of system actions when people find them unintelligible. More specifically, they are unable to engage in clarification and repair dialogues and lack the capacity to dynamically revise their model of their own actions in response. Different users, and different user groups all interpret actions in different ways s intelligent systems need the ability to acquire and adapt to the user’s language sometimes in the course of a single interaction (Healey, 2008)

This problem of machine accountability is aggravated by the increasing availability of complex data sets and the use of non-linear approaches to data reduction and machine learning techniques, such as.SVM’s and multi-layer neural nets, which produce models that are not directly interpretable even by their designers (Vellido et.al. 2012) Like humans, machines using these techniques will sometimes be unable, even in principle, to give a direct account of the computations that produce particular results. Overcoming this limitation, we propose, will involve giving machines the ability to develop, in effect, human-like ‘intentional’ theories of their own behaviours that are primarily tested and revised through interactions with users.

Accountable Intelligence:

The long-term success and acceptability of intelligent machines will depend on their accountability. As the behaviour of systems becomes more complex so will the expectation that we can engage with them about the reasons for their actions. This applies as much to domestic thermostats as it does to personal digital assistants. Currently, the most sophisticated interactive systems such as SIRI, Cortana and Alexa fall far short of providing the kind of interactive flexibility that humans are capable of (Luger and Sellen, 2016).

This issue is more than a matter of usability. If intelligent systems are to act as proxy for human agency they will also need to be able to proxy for human accountability. For example, legal accountability for medical decision making rests with the physician. If they use decision support systems they are expected to be able to assess and override the information they provide (Berner, 2002). In this context the ability of the machine to produce credible and reliable explanations of its own behaviours is critical.

Key Challenges:

The challenges involved in building accountable intelligences are considerable but they are mitigated to some extent by what is already known about the mechanisms that underpin human accountability. Here we identify some key areas in which models of human interaction could usefully inform new research effort in accountable intelligence:

- Common ground: the ability to model the accumulation of shared conversational contexts that are specific to particular interactions and particular users.

- Hybrid Architectures: the ability to represent and reason about system computations and behaviours in a language that is intelligible to users.

- Collaborative Adaptation: the ability to make real-time revisions to the language of system computations and behaviours in response to user feedback.

- Resource Plasticity: the ability to recognise and deploy ad-hoc extra-linguistic resources e.g. gestures, objects and environment. as part of a structured interaction.

- Moral Responsibility: the ability to represent and reason about the social norms relevant to machine actions and to respond to user's signals of possible deviations from relevant norms.
  1.
- Conceptual Change: collaborative adaptation may require changes in the *language* of the system's models, as well as changes in the beliefs represented in those models.

References:

E. S. Berner, Ethical and legal issues in the use of clinical decision support systems, *Journal of Healthcare Information Management*, 16(4):34-37, 2002.

Bundy, A. (in prep) Smart Machines are Not a Threat to Humanity.

Bellotti, V. and Edwards, K. (2001) Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human–Computer Interaction* Vol. 16, Iss. 2-4, 2001.

Dennett, D. (1987) *The Intentional Stance*. MIT Press.

Garfinkel, H. (1984) Studies in Ethnomethodology. Polity Press/ Blackwell Publishers, Oxford, UK. First published in USA, 1967.

Healey P.G.T. (2008) "Interactive Misalignment: The Role of Repair in the Development of Group Sub-languages" in Cooper R. and and Kempson R. (eds) *Language in Flux: Relating Dialogue Coordination to Language Variation, Change and Evolution*. Palgrave-McMillan. pp 13-39. ISBN: 978-10904987-96-3.

Luger, E., & Sellen, A. (2016, May). Like Having a Really Bad PA: The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5286-5297). ACM.

Osman, M. (2010) "Controlling Uncertainty: A Review of Human Behavior in Complex Dynamic Environments" Psychological Bulletin, v.136 (1) pp. 65-86.

Vellido, A., Martın-Guerrero, J.D. and Paulo J.G. (2012) Making machine learning models interpretable. LisboaESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), pp. 25-27 April 2012.