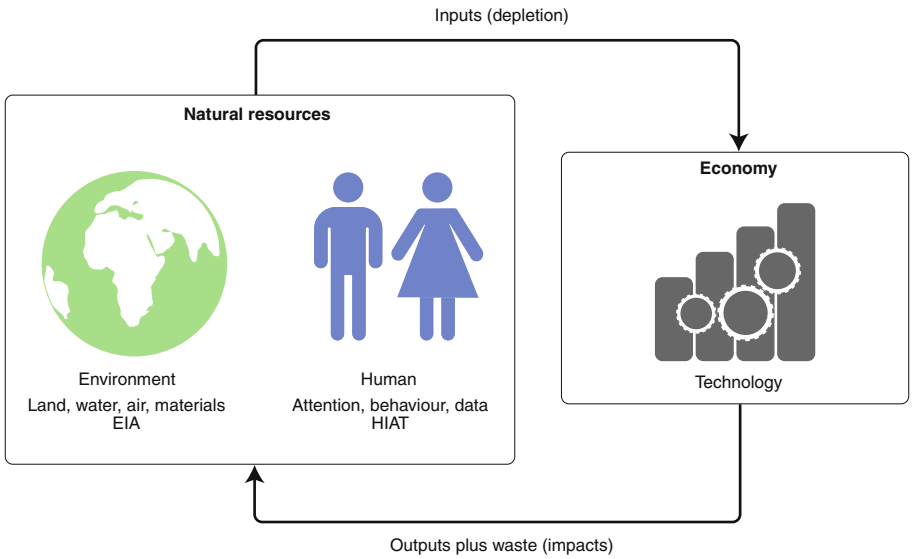
Advancing impact assessment for intelligent systems

With the rise of AI technologies in society, we need a human impact assessment for technology.

Rafael A. Calvo, Dorian Peters and Stephen Cave

It is the 50th anniversary of the environmental impact assessment (EIA), a significant step towards making engineering more socially responsible. But a growing number of policymakers are now voicing the need for an approach to assess the human and social impacts of intelligent systems. We discuss how the EIA provides a partial blueprint for what we call a human impact assessment for technology (HIAT), and how more recent algorithmic and data protection impact assessment initiatives can contribute. We also discuss how ethical frameworks for such a human impact assessment could draw on recently established artificial intelligence (AI) ethics principles. We argue that this approach will help build trust in an industry facing increasing criticism and scrutiny.



Human and environment as subjects of resource extraction and impact.

Outputs and costs of data-driven industries

In 1869, amid the enthusiasm of the first industrial revolution, T. H. Huxley called on readers of the first issue of *Nature* to rejoice in the progress of the previous 50 years[1]. In 1969, exactly 100 years later, the environmental costs of that progress had become salient enough to prompt radical changes to policy through formal environmental assessment and regulation.

Speeding forward another 50 years, we now face a fourth industrial revolution and our relationship with technology is again in flux. While we benefit from this progress, we are also witnessing its ethical, psychological and social costs. Some of the by-products of our systems include manipulation of emotions[2], attention[3] and voting behaviours at scale[4], as well as interactions designed for deception and coercion[5]. AI systems can bias legal, educational and employment decisions[6] and have unintended negative impacts on health and wellbeing[7].

But we can learn from the history of EIA for improving the impacts of AI. Like environmental impact, the human impact of intelligent technologies is difficult to model, with consequences that are hard to predict. It also requires a framework for strong multidisciplinary collaboration, and multi-sector engagement involving industry, regulators and the general public. Moreover, the 50-year history of experience with EIA can teach us to anticipate likely issues including disagreement about measures, the need for an evolving process, and conflicts between industry and regulation.

Traditional EIA evaluates the effects of human intervention on the biophysical environment by considering the intended (products) and unintended (waste) consequences of industry. The process allows for the identification of resource costs (for example, depletion of fossil fuels) and pollutants (for example, chemical run-off) so these can be weighed against potential benefits. New, data-driven industries differ from conventional engineering projects as human activity itself is considered the resource as well as the product, while individual or societal ill-being is a potential 'waste' effect. Engineers and other stakeholders such as policymakers need an assessment framework that takes this into account.

While the natural environment is an involuntary recruit to human industry, and has no voice of its own, data industries rely on human resources to generate their data, and it is a general democratic principle that humans are given a say with regard to actions that will affect them. As such, we will need to improve the ways human needs and values are integrated into design processes with respect to AI. Measuring impact will help make these values explicit and thus provide a platform for defending them.

Distinguishing HIAT

While the EIA has inspired a number of human- and society-focused processes such as health impact assessment[8] and social impact assessment[9], none is sufficient for addressing the far-reaching and unprecedented effects of intelligent systems due to a number of unique characteristics. Specifically, an effective HIAT will need to manage the following:

- Dynamic systems. While traditional impact assessment approaches evaluate interventions that are relatively static once in place, intelligent systems can

change rapidly. With few physical limitations, software is modified easily, remotely and, in addition to regular updates, intelligent systems are continually gathering new data, learning and reprogramming themselves.
- Scale. While traditional approaches assess impact within communities or regions with geopolitical boundaries, software has no physical boundaries and, as a society, we have witnessed impacts crossing geographic, political, legal and cultural borders.
- Responsiveness. While the emphasis in traditional impact assessment is on the anticipation of impact, and monitoring to check for compliance, the scale and malleability of software impact make it harder to predict accurately. Therefore, while anticipation is still essential, an effective HIAT process will have to lean more heavily on ongoing evaluation and improvement in response.
- Humans as resource. As noted above, while traditional EIA centres on the natural environment as the source of resource extraction, new intelligent systems centre on humans. Many of these systems mine human data and extract human attention, using these to change human behaviour. Such an approach has novel social and philosophical implications that need to be considered.
- Practice and training. While traditional impact assessment is standard to physical engineering, policy and government practice, software designers and engineers currently have no impact assessment built into their processes or training. This must change. To make this change possible, a HIAT must be designed with consideration for the iterative and agile processes common in these fields.

A handful of recent efforts have already contributed to impact assessment processes for specific areas related to AI. These include data protection impact assessments (DPIAs) required by EU regulations, and algorithmic impact assessments (AIAs) focusing on automated decision-making in government (for example, for policing, resource-allocation and so on). For example, the AI Now Institute outlines an AIA process for government procurement of automated decision systems[10] and the Canadian government provides a scorecard tool for evaluating such systems. The AIA framework is intended to help public agencies to critically assess automated systems that impact justice and fair distribution. AI Now states that the AIA is "designed to support affected communities and stakeholders as they seek to assess the claims made about these systems, and to determine where — or if — their use is acceptable." As such, it is at the time of procurement that agencies conduct an AIA. In contrast, the HIAT aims to become part of the design and development process from early stages, and to involve technology-makers themselves.

A HIAT would introduce a framework large enough in scope to incorporate all uses of intelligent technologies while being sensitive to the factors unique to these systems. In the next section we briefly outline some practical approaches towards achieving this.

Defining and measuring impact

A HIAT should aim to predict and evaluate the impact that new digital technologies have on all stakeholders. It would acknowledge that these systems have psycho-social impacts on individuals and society, and that some of these may be negative for health, wellbeing and values such as privacy, autonomy and democracy.

While the EIA deals with physical and observable measures, the most salient effects of intelligent systems are often on the subjective experience of human beings. Social scientists have established means for measuring this kind of impact. For example, the World Happiness Report published by the UN Sustainable Development Solutions Network[11] measures population-wide wellbeing across countries using measures from psychology and behavioural economics. Human–computer interaction researchers also use a combination of qualitative and quantitative methods from psychology for the evaluation of software systems with respect to psychological experience[12]. Some of these draw on measures of autonomy and wellbeing developed over several decades in psychology research[13].

For a HIAT report, qualitative and quantitative measures could be given for a range of categories determined by the ethical frameworks for intelligent systems already in development, such as the IEEE Global Initiative on Ethical AI. This large-scale international initiative is producing a set of industry standards akin to ISO 14001:2015 (a management system for environmental responsibility). The IEEE specifications will include assurance procedures for technical issues affecting privacy, security, psychological wellbeing and other ethical concerns. While these specifications promise detailed technical standards, other ethical frameworks, such as the Ethics Guidelines for Trustworthy AI[14] may provide more workable summary structures for an assessment and report. While there are many AI ethics frameworks available, recent analysis shows that they share common value structures and equate to other ethical frameworks in health[15], providing helpful corroboration and alignment with other forms of professional practice and regulation.

In short, we would make the following practical recommendations for developing a first version of a HIAT:

- Use an existing ethical technology framework such as the European Union's Ethics Guidelines for Trustworthy AI as a foundation to guide assessment and report structure.
- Employ social science methods for measuring impact against the framework — for example, measures of autonomy and wellbeing developed in psychology.
- Include existing assessment methods and tools developed for specific technologies such as automated decision systems (AIA) and data privacy (DPIA) as appropriate to the context.
- Require compliance with related technical standards, such as IEEE specifications as they emerge.
- Gather input from all stakeholders, including a diverse representative sample of end-users, in order to inform the assessment.

Building trust

While pressure for greater regulation of intelligent systems is mounting, the dominant argument against is that it slows innovation agility. Such concerns apply to all industries, but agility has been of special concern in the software industry, which has until recently favoured the 'move fast and break things' motto. But as this industry comes under increasing criticism and scrutiny, it is realizing the importance of systems that build trust. Other industries know that values such as quality assurance, safety, security and traceability are essential to this. Impact assessments are an important tool for embedding and validating these values and have been successfully used in many industries including mining, agriculture, civil engineering and industrial engineering. Other sectors too, such as pharmaceuticals, are accustomed to innovating within strong regulatory environments, and there would be little trust in their products without this framework. As AI matures, we need frameworks such as HIAT to give citizens confidence that this powerful new technology will be broadly beneficial to all. ❐

Rafael A. Calvo[1,2*], Dorian Peters[1,2] and Stephen Cave[2]

[1]Dyson School of Design Engineering, Imperial College London, London, UK. [2]Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK.

**e-mail: r.calvo@imperial.ac.uk*

Published online: 3 February 2020
https://doi.org/10.1038/s42256-020-0151-z

References

1. Huxley, T. H. *Nature* **1**, 9–11 (1869).
2. Kramer, A. D. I., Guillory, J. E. & Hancock, J. T. *Proc. Natl Acad. Sci. USA* **111**, 8788–8790 (2014).
3. Wu, T. *The Attention Merchants: The Epic Scramble to Get Inside Our Heads* (Vintage, 2017).
4. Bond, R. M. et al. *Nature* **489**, 295–298 (2012).
5. Burr, C., Cristianini, N. & Ladyman, J. *Minds Mach.* **28**, 735–774 (2018).
6. Crawford, K. & Calo, R. *Nat. News* **538**, 311–313 (2016).
7. Kross, E. et al. *PLOS ONE* **8**, e69841 (2013).
8. Kemm, J., Parry, J., Palmer, S. & Palmer, S. R. (eds.) *Health Impact Assessment* (Oxford Univ. Press, 2004).
9. Becker, H. A. *Eur. J. Oper. Res.* **128**, 311–321 (2001).
10. Reisman, D., Schultz, J., Crawford, K. & Whittaker, M. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability* (AI Now Institute, 2018).
11. Sachs, J., Becchetti, L. & Annett, A. *World Happiness Report 2016* Vol. 2 (UN Sustainable Development Solutions Network, 2016).
12. Peters, D., Calvo, R. A. & Ryan, R. M. *Front. Psychol.* **9**, 797 (2018).
13. Deci, E. L. & Ryan, R. M. *Psychol. Inq.* **11**, 227–268 (2000).
14. *Ethics Guidelines for Trustworthy AI* (European Commission, 2019).
15. Floridi, L. et al. *Minds Mach.* **28**, 689–707 (2018).

Competing interests

The authors declare no competing interests.