Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Human Autonomy in AI Systems

This input paper was prepared by Rafael Calvo, Dorian Peters and Richard Ryan

Suggested Citation

Calvo, R, Peters, D and Ryan, R (2018). Human Autonomy in AI Systems. Input paper for the Horizon Scanning Project “The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing” on behalf of the Australian Council of Learned Academies, www.acola.org.

Supporting human autonomy in AI systems

Rafael A Calvo[(1)], Dorian Peters[(1)], Richard M. Ryan[(2)]

[(1)]School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW, Australia

[(2)]Institute for Positive Psychology and Education, Australian Catholic University, Sydney, NSW, Australia

Abstract

Since time immemorial humans have sought “autonomy”, the feeling that one is acting in accordance with their goals and values. Today, the meaning of, and the path towards autonomy are being reassessed in the context of the growing impact of artificial intelligence (AI). This impact is sometimes seen as a dialectical relationship where humans are in constant conflict with machines, or as a synergistic one, in which the two need to collaborate efficiently. Both perspectives fail to consider the way a technology is experienced by its users. We posit that, if AI is to support human wellbeing, developers need to consider the impact of AI on users’ experience of autonomy. We describe a model for the design of computer systems that takes autonomy into account as a basic psychological need, and we provide recommendations for designers and policy makers.

Introduction

The public’s concerns regarding AI are often built on the idea that humans will increasingly concede autonomy to what is perceived (and often hyped) as a robotic uprising; technologies that will enslave individuals. This account sees technologies as contributing to an alienated population that lacks agency and is increasingly detached from the outcomes of their work. In a different account, most commonly used in engineering, the relationships between humans and machines tend to be explored in the context of maximizing productivity and safety. In this account the designer works as a choreographer, planning the interactions between humans and AI systems. If we want AI to maximise human wellbeing and economic benefits across the economic spectrum we need to go beyond these accounts.

These two accounts about the impact of technology derive from the potentially opposing goals facing designers, with one focussed on technological prowess and the other on users’ wellness and productivity. Although seemingly opposed, recent evidence suggests that they need not invariably be. Interaction and system designers in what is often termed human-computer interaction (HCI) research are finding that both sets of goals are connected to the way stakeholders psychologically

experience their relationships with various technologies/AIs. Specifically technologies vary not only in what they can accomplish, but how they impact user experience in adaption, use, impact and wellness, HCI researchers, working with psychologists, are developing methods and instruments to assess how any particular design is experienced by its users. For example, *self-determination theory* (SDT; Ryan and Deci 2017) was used as the foundation for an HCI model for "Motivation, Engagement and Thriving in User Experience" (METUX) (Peters, Calvo and Ryan, 2018). This model suggests that design elements that enhance users' experience of autonomy can also enhance that positive impact of design on wellness and effectiveness of use.

But the impact of AI is broader than just this experiential interface, and involves ethics, social, cultural and economic issues, all of which may affect the wellbeing of technology users. Policy making brings these other considerations into the picture. In fact, the process of designing computing systems has many similarities with policy making: it starts with a design brief that describes the purpose and desired outcome for the system – similar to the purpose statement in a policy document. In both, there are steps for formulating the solution, then implementation, evaluation and termination. Government and corporate policies can themselves be used to prompt designers to formulate designs that consider the impact that they will have on users on all these fronts. This is not new: product safety guidelines and regulations have radically improved the way designers consider the physical health and safety of products. Given the impact of AI on psychological health, new regulations will need to consider how AI is used in ways that respect the psychological health of individuals.

This paper describes a model for evidence-based design that could drive an AI product policy. The empirical model is based on SDT, a body of psychological research that has already influenced education and health systems. SDT holds that technologies differentially support basic human psychological needs of human autonomy, sense of competence and relatedness. In this article we will focus only on autonomy, and how users' experience of autonomy should be a central consideration in order to develop effective and psychologically beneficial AI. SDT also suggests criteria for evaluating the impact of technologies on people's basic psychological needs, which can inform larger polices and practices.

Background

Engineers and technology designers have traditionally sought to maximize productivity. This is particularly true for software systems that are unlikely to cause physical harm to the users. But the psychological impact of new technologies is now obvious and has become a design imperative. In the last five years researchers have developed new design methods to support psychological wellbeing. Research in this area has been called Positive Computing (Calvo and Peters, 2014), Experience Design (Hassenzahl, 2010), Positive Design (Desmet and Pohlmeyer, 2013). They often build on psychological theories that provide models that can be used to understand ways in which using technology influences the (1) affective quality, (2) engagement/actualization, and (3) connectedness of experience.

On such theory is SDT (Ryan and Deci, 2000, 2017), which examines the factors that promote sustained motivation and wellbeing. The theory has gathered one of the

largest bodies of empirical evidence in psychology and identifies a small set of basic psychological needs deemed *essential* to people's self-motivation and psychological wellbeing. Furthermore, it has shown how environments that neglect or frustrate these needs are associated with ill-being and distress. These basic needs are:

- **Autonomy** (feeling agency, acting in accordance with one's goals and values),
- **Competence** (feeling able and effective),
- **Relatedness** (feeling connected to others, a sense of belonging).

In this article we only focus on the individual's need for autonomy, but note that aiming to support all three is important for optimal functioning. For example, some authors (e.g. Carr, 2015) have raised concerns that AI systems contribute to losing competencies, such as wayfinding without the support of a GPS enabled map. Others (eg. Turkle, 2017) have suggested that AI enabled technologies can contribute to loneliness and the lack of meaningful relationships. We chose to focus on autonomy in this article since it is often forgotten by engineers (who focus on the machine) and by policy makers who do not always appreciate the trade-offs that designers must make.

Here we refer to an autonomous person as one that has a sense of willingness, endorsement, and/or volition in acting (Ryan and Deci, 2017). This is not the same as doing things independently or being in control; rather it means acting autonomously and in accordance with their personal goals and values. Individuals often relinquish control or embrace interdependence on their own volition.

Within engineering, the vast majority of research has focused on the design of autonomous *systems*, particularly robots and vehicles, rather than on supporting autonomous *humans* (Baldassarre et al., 2014). More recently however, the Institute of Electrical and Electronics Engineers (IEEE) has developed a charter of ethical guidelines for the design of autonomous systems that places *human* autonomy and wellbeing at center-stage (Chatila et al., 2017). Such focus has been common in some areas of design research (Calvo et.al. 2014) including Value-Sensitive Design. For example, in the context of software systems, Friedman (1996) identified system capability and complexity, misrepresentation, and fluidity as key design factors that can support or hinder user autonomy. Friedman only considered the impact of the system's interface and not the broader impact on other aspects of a person's life. This is a limitation that the METUX model, discussed next, addresses.

Motivation, Engagement and Thriving in User Experience" (METUX)

The METUX model introduced by Peters, Calvo and Ryan (2018) draws on research that uses SDT to better understand psychological processes in the use of technologies such as video games and, more generally, empirical research in workplaces, health, and education contexts (see Ryan and Deci, 2017 for a review). The METUX model describes and predicts how a technology affects motivation, engagement, and wellbeing based on how the technology satisfies or thwarts psychological needs.

The model introduces six separable "spheres of experience": Adoption, four that address individuals and can be designed for: Interface, Task, Behaviour, Life; and

Society. A full description of these is out of scope but examples are provided in (Peters, Calvo and Ryan, 2018) and in the next section. Broadly speaking the Interface is the software itself – what users interact with. A Task refers to an activity or action created by the technology: finding a paper or author in Google Scholar, or tracking steps with a smart watch. The sphere of Behaviours refers to those enabled or enhanced by the technology: exercising (by a health app), or reviewing of academic literature (Google Scholar). The previous spheres refer to momentary need satisfaction, while the Life sphere refers to the longer timespan.

The separation between these spheres is useful to designers, but we note that they overlap and interrelate so, in some contexts, they need to be adapted. The aim is to provide a way of organizing thinking and evaluation in a way that can address contradictory parallel effects (i.e., a technology can support psychological needs at one level while undermining them at another). Figure 1 provides a visual representation of the model.

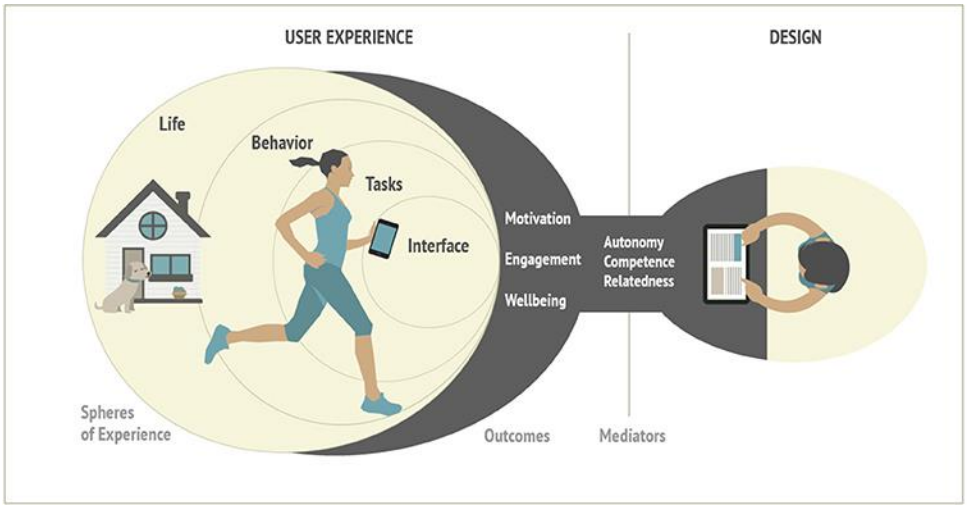

Figure 1: METUX model and its use in technology design

Example: personal profiling using AI

Consider the design of an automated personal profiling tool such as Google Scholar (that aggregates all the publications of an academic), or a system that automatically identifies a person showing signs of mental illness in their social media posts (Calvo, et al 2017). These systems collect large amounts of textual data and use AI in natural language processing to extract information. The result is either a structured summary of an academic's research, or a label that identifies an individual as mentally-ill or not (often with a more specific diagnosis of the illness). In both cases the automated profiles are high stakes for individuals. For example, Google Scholar can be used in hiring and promotion decisions, or a misdiagnosis may result in an automated triage system can leave someone without the help they need.

The design of these system can be based solely on productivity and automation. Such a design brief would lead designers to decide that the individual being profiled does not even need to know he is being profiled, and no human needs to be involved in the profiling process. On the other hand, a design brief that takes into account the experience of the user would consider involving the individual, maybe allowing her to

customize the results, decide what is done with them (do they become public?, do they trigger actions from others?). The fact that the data about the individual is available should not imply that design decisions on what to do with it can be taken lightly.

Table 1 provides a summary description of some of the design questions (and how they have been addressed) for these two technologies. The Google Scholar descriptions are based on direct usage of the system by the authors. The description of the Mental Health Triage system is based on the system described by Calvo et. al (2017) currently used by ReachOut Australia to support over 1.8M Australian users. This system uses sophisticated AI with natural language processing tools described by Altszyler et.al (2018). The METUX model provides instruments (i.e. questionnaires) that could be used to measure the impact of different design choices on users, but the impact of designs in these two scenarios have not been empirically evaluated.

Table 1: Spheres of experience for two AI-based personal profiling tools

Sphere of Experience	Google Scholar	Mental Health Triage
Adoption	Currently user decides to make profile publicly available	Currently community expects moderation and the triage is designed to help user moderators
Interface *Is direct interaction possible and how does it affect needs satisfaction?*	currently users can only enter basic personal information and photograph)	Only moderators view this information. Moderators do not control how the interface looks but can prioritize and redirect cases
Tasks *What are the technology specific tasks? How do they support needs satisfaction?*	Few tasks for the profile owner	No tasks available to users
Behaviour *How does the technology improve needs satisfaction for the behaviour the technology supports?*	Academic profiling is complex, and is likely to have an impact on self-concept.	No behaviours expected from users.
Life *How does the technology influence needs satisfaction in life overall?*	Google scholar is an efficient instrument for comparing academic outcomes. It is not clear what the overall psychological consequences are.	Triage system helps users receive better and more timely care

Organisations who produce AI technologies could be required to consider the consequences of their designs on the wellbeing of an individual. The design process could include a phase where the impact of the AI on the individual's sense of autonomy is measured. It is important to note that the design does not always have to be a trade-off between productivity and the individual. Over time, design patterns that optimize productivity together with the psychological needs of the individual will be developed. In the meantime, organisations that introduce new technologies can evaluate the impact they are having, and make informed procurement decisions.

Ethics in the design of AI systems.

Different ethical theories could be used in design. An ethical approach to the design of AI systems is one that promotes the "good life", that is, a life of virtue in Aristotelian ethics. According to a deontic approach it requires following rules no matter the consequences, and according to a utilitarian approach it requires maximizing wellbeing of the largest number of people possible (independent of the means). Arguably these three different ethical perspectives provide different guidelines for designers and policy makers.

But above all, these ethical theories value (or at least are not opposed to) evidenced-based approaches to promoting wellbeing. The empirical evidence is clear: promoting individuals' agency, competence, and sense of relatedness (in health, workplaces, and education), lead to experiences of psychological wellbeing and are therefore the most ethical choices (see Ryan and Deci 2017 for a review). In fact, since they also promote engagement and productivity they are arguably beneficial to other stakeholders and society at large.

Australia has a proud tradition of fair practices in workplaces, education and health, all of which are now being impacted by AI. Australia also has world-class research in the area of AI. Taken together: a supportive socio-political, technological and academic environment puts Australia in an enviable position to lead the world in the ethical design of new technologies.

Acknowledgements
RC is supported by an Australian Research Council Future Fellowship (FT140100824). The authors acknowledge Emma Bradshaw for her feedback on the manuscript.

References

Altszyler, E., Berenstein, A. J., Milne, D., Calvo, R. A., & Slezak, D. F. (2018). Using contextual information for automatic triage of posts in a peer-support forum. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*(pp. 57-68).

Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M., and Barto, A. (2014). Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. *Front. Psychol.* 5:985. doi: 10.3389/fpsyg.2014.00985

Carr, N. (2015). The glass cage: Where automation is taking us. Random House.

Calvo, RA, Peters, D., Johnson, D, Rogers Y. (2014) “Autonomy in Technology Design” CHI ’14 Extended Abstracts on Human Factors in Computing Systems Pages 37-40. ACM, 2014

Calvo, R. A., Hussain, M. S., Milne, D., Nordbo, K., Hickie, I., & Danckwerts, P. (2017). Augmenting online mental health support services. In Gaming and Technology Addiction: Breakthroughs in Research and Practice (pp. 264-285). IGI Global.

Calvo, R., and Peters, D. (2014). *Positive Computing: Technology for Wellbeing and Human Potential*. Cambridge, MA: MIT Press.

Calvo, RA, D. Milne, SM Hussain, H Christensen “(2017) “Natural language processing in mental health applications using non-clinical texts”. *Natural Language Engineering* , 23(5), 649-685

Chatila, R., Firth-Butterflied, K., Havens, J. C., and Karachalios, K. (2017). The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems [standards]. *IEEE Robot. Automat. Mag.* 24, 110–110. doi: 10.1109/MRA.2017.2670225

Desmet, P. M. A., and Pohlmeyer, A. E. (2013). Positive design: An introduction to design for subjective well-being. *Int. J. Design* 7, 5–19.

Friedman, B. (1996). Value-sensitive design. *Interactions* 3, 16–23. doi: 10.1145/242485.242493

Hassenzahl, M. (2010). Experience design: technology for all the right reasons. *Synth. Lect. Hum. Center. Informat.* 3, 1–95. doi: 10.2200/S00261ED1V01Y201003HCI008

Peters, D, Calvo, RA, Ryan, RM “Designing for Motivation, Engagement and Wellbeing in Digital Experience” *Frontiers in Psychology* – Human Media Interaction. Vol 9. pp 797.

Ryan, R. M., and Deci, E. L. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066X.55.1.68

Ryan, R. M., and Deci, E. L. (2017). *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York, NY: Guilford Press.

Turkle, S. (2017). Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books.