Human Values and AI,
A Values in Computing (ViC) Workshop.

Maria Angela Ferrario, Emily Winter, Stephen Forshaw.
School of Computing and Communications, Lancaster University, UK.

Abstract: Artificial Intelligence (AI) has seen a massive and rapid development in the past twenty years. With such accelerating advances, concerns around the undesirable and unpredictable impact that AI may have on society are mounting. In response to such concerns, leading AI thinkers and practitioners have drafted a set of principles - the *Asilomar AI Principles* - for Beneficial AI, one that would benefit humanity instead of causing it harm. Underpinning these principles is the perceived importance for AI to be *aligned* to human values and promote the 'common good'. We argue that efforts from leading AI thinkers must be supported by constructive critique, dialogue and informed scrutiny from different constituencies asking questions such as: what and whose values? What does 'common good' mean, and to whom? The aim of this workshop is to take a deep dive into human values, examine how they work, and what structures they may exhibit. Specifically, our twofold objective is to capture the diversity of meanings for each value and their interrelationships in the context of AI. We will do so both systematically and creatively using tools and techniques developed as part of the Values in Computing (ViC) research. In practice, we will engage in a small set of facilitated group activities designed to explore the *Asilomar AI Principles* in the context of a broader values theoretical framework briefly outlined in this paper.

Keywords: values in computing, human values, artificial intelligence, ethics.

Background

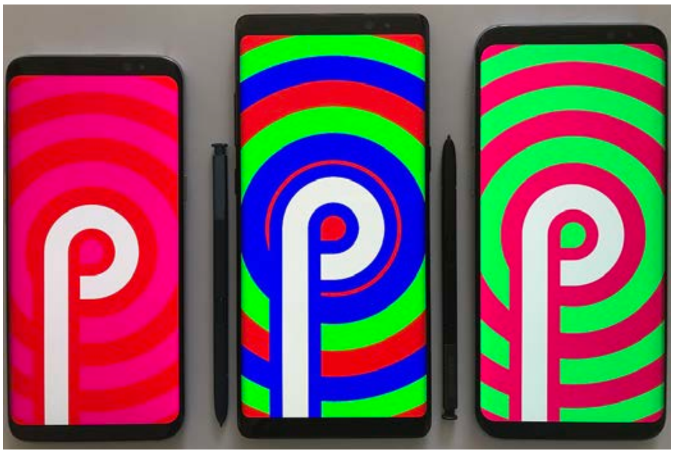
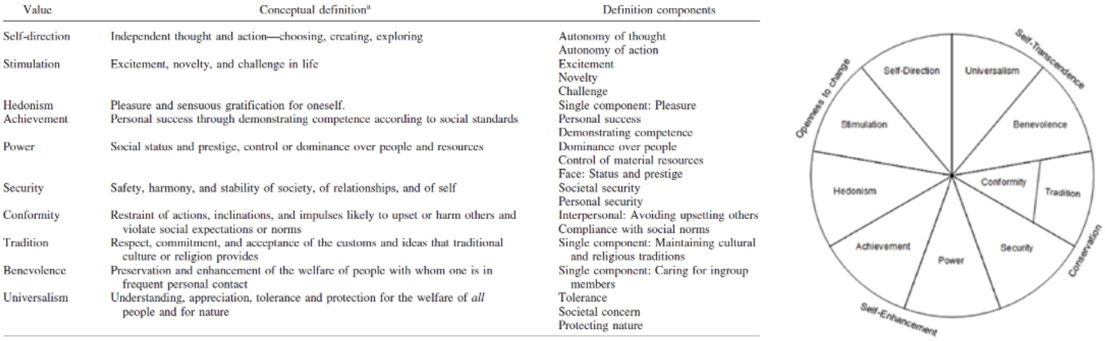


Figure 1. The latest Android upgrade, Android 9 Pie, contains Smart Linkify, an API that uses machine learning for entity recognition to add clickable links when certain entities (e.g. street addresses) are detected in text (Samar 2018). Photo: https://commons.wikimedia.org/wiki/File:Android_9.0.jpg CC BY 2.5 license.

Andrew Ng, Google Brain's co-founder and one of today's most influential AI figures, has likened AI to electricity for its transformative power and pervasiveness (Lynch 2017). AI's rapid advances and range of application domains - from 'mundane' apps (Figure 1), to cancer screening (Ting 2018) and military intelligence (Suchman 2017) - have also raised questions regarding the desirability of such advances for society. A constructive and informed discussion around AI is not always easy given that AI is often shrouded in media hype and technical jargon. There are several definitions of AI. We find Lucy Suchman's particularly helpful: "AI is the field of study devoted to developing computational technologies that automate aspects of human activity *conventionally understood* to require intelligence" (Suchman 2018). We argue that the expression 'conventionally understood' (our *emphasis*) is key. In the last twenty years, the AI field of study has been focused on the "construction of intelligent agents - systems that perceive and act in some environment. In this context, the *criterion for intelligence* is related to statistical and economic notions of rationality - the ability to make good decisions, plans, or inferences" (Russell 2015). The criterion for intelligence (our *emphasis*) is here linked to rational thinking as defined by statistics and economics – e.g. in terms of *utility*, and *performance* (Russell 2016).

AI / human values alignment?



Value	Conceptual definition[a]	Definition components
Self-direction	Independent thought and action—choosing, creating, exploring	Autonomy of thought Autonomy of action
Stimulation	Excitement, novelty, and challenge in life	Excitement Novelty Challenge
Hedonism	Pleasure and sensuous gratification for oneself.	Single component: Pleasure
Achievement	Personal success through demonstrating competence according to social standards	Personal success Demonstrating competence
Power	Social status and prestige, control or dominance over people and resources	Dominance over people Control of material resources Face: Status and prestige
Security	Safety, harmony, and stability of society, of relationships, and of self	Societal security Personal security
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms	Interpersonal: Avoiding upsetting others Compliance with social norms
Tradition	Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provides	Single component: Maintaining cultural and religious traditions
Benevolence	Preservation and enhancement of the welfare of people with whom one is in frequent personal contact	Single component: Caring for ingroup members
Universalism	Understanding, appreciation, tolerance and protection for the welfare of *all* people and for nature	Tolerance Societal concern Protecting nature

Figure 2. Schwartz's Values Model; left: conceptual definitions of 10 basic values (Schwartz 2012); right: visual representation of the model capturing observed interrelationships between values (Schwartz 1992).

If we look at Schwartz's values model (Schwartz 1992, 2012 – Figure 2), we notice that concepts such as utility and performance, and in general the concepts underpinning the fields of economics and statistics, tend to emphasise a values subset of the model. Specifically, they link to the subset that includes the values of *achievement* (e.g. quantifiable success –reward vs punishment in reinforced learning (Russell 2016)), *power* (e.g. over resources), and *security*. These are indeed human values, but we must be mindful of the values 'bias', which may be inherent to AI, when calls for values alignment in AI are made (Arnold 2017, Hadfield 2016, Riedl 2016). Schwartz's values model is one of the most extensively and empirically investigated values models to date. It is not free from limitations (Maio 2010), but it is a useful starting point for investigation.

The Asilomar AI Principles

Top AI researchers have long reflected and written on ethical and existential concerns around AI (Bostrom 2005, 2017). In 2017, leading "AI researchers from academia and industry, thought leaders in economics, law, ethics, and philosophy" came together for a five-day conference on Beneficial AI (Future of Life Institute 2017). One the key outcomes of that gathering was an agreement on a set of principles - the *Asilomar AI Principles* – to provide guidance to the development of beneficial AI (Asilomar AI Principle 2017).

Underpinning them all is a genuine concern for human values, but one that we find quite broad and undefined. Principle 10, for example, focuses on "Value Alignment" between autonomous systems and humans; Principle 23 focuses on "Common Good" and states "Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity". Whose and what ethical ideals?

The limitation of 'Ethics'

Outside the AI-specific domain, a large body of computing research also exists stressing the importance of building technologies that embody human values. This work spans from Values Sensitive Design (Friedman 2006, 2017) to ethical computing (Van de Hoven 2012).

However, also within this large body of research, the tendency seem to focus primarily on values with ethical import (LeDantec 2009), which are a subset of a much broader human values set that includes wealth, prestige, and power (Figure 2). We find this problematic because these latter are human values too and they *do* already drive digital technology production and shape their use (Ferrario 2016, 2017). Failing to consider this, risks not capturing the interdependences, tensions, agencies and relations between values.

In addition, values such as ‘freedom’ and ‘public good’ are often poorly articulated, thus running the risk of becoming *cultural truisms,* “beliefs that are widely shared and rarely questioned” (Maio 1998). Crucially, research (Hanel 2018) has also found that the same value can have different meanings for different people and cultures; and this not only at an abstract personal level, but also at a behavioural, or instantiation, level (Maio 2010)

Towards a better understanding of human values

This poor articulation of values can have far-reaching implications on our personal and social lives. For example, ‘freedom’ may be called upon to protect the fundamental human rights of unlawfully imprisoned civilians, but also to ‘free’ one country from the presence of law-abiding immigrants. The psychologist Greg Maio has recently called for a more scientific understanding of values (Maio 2018), whilst his previous work (Maio 2010) gives poignant examples of the variety of meanings that the same value may have in political discourse.

Our research aims to help address the call for a more ‘scientific’ understanding of values in the computing domain (e.g. systematic, empirically based, reproducible). We do so by drawing principles and techniques from a variety of disciplines and in particular from those that have taken an empirical approach to the study of human values (Schwartz 1992, 2012; Maio 2010). This is not research for the few. Rather, it requires the participation and deliberation of many. In this workshop, we combine such methods with creative design thinking techniques (Forshaw 2012, Newman 2015) to create both a base of common understanding around AI and systematic reasoning around human values.

Our ultimate goal may not be to program values-driven intelligent agents, but to support the next generation of educators and computing professionals with “the deliberative, technical, and critical skills necessary to tell the difference between what is worth pursuing from what is potentially harmful to self and society” (ViC team 2018).

Theoretical Framework, Tools & Approach

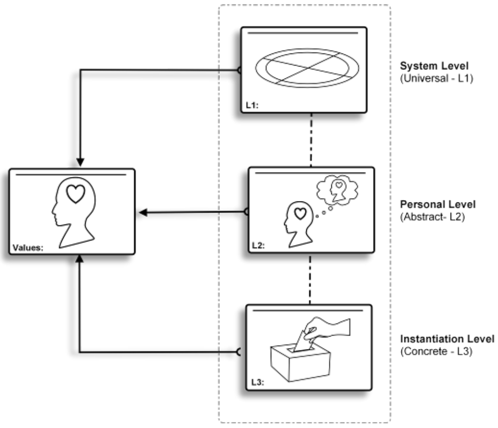


Figure 3. Values as mental representations to be studied on three levels: system (L1), personal (L2), and instantiation (L3).

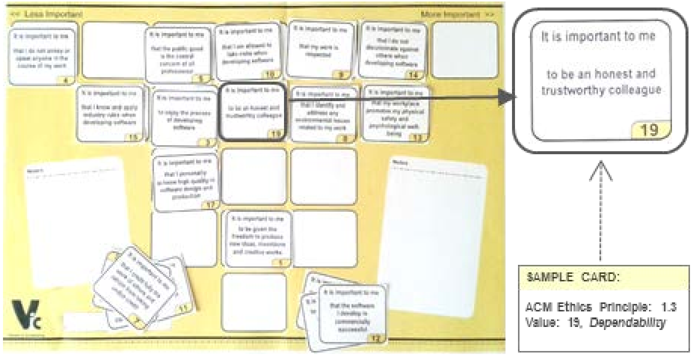
We considers values as mental representations to be investigated on three levels: at a system (L1), personal, (L2), and instantiation level (L3) – Figure 3. This 3-level theoretical framework is based on an established body of work from experimental psychology, which draws from Schwartz’s values model (Schwartz 1992) and Maio’s work (Maio 2010).

L1 concerns the existence of a values model (Schwartz 1992, 2012) where the mental representations of human values are found to occur according to certain observed patterns. For example, people who value social power highly have been found to value equality less.

L2 relates to the abstract representation that an individual has for each value and the variety of meanings associated to it. For example, the mental representation of ‘freedom’ held members of different political parties.

L3 relates to the way values drive or at least influence actions. The instantiation level is the most difficult to study. One can never assume a direct link between actions and values, and the same value may drive different actions. For example, for some caring for the environment means recycling waste, for another is marching against shale gas extraction.

Tools



<< Less Important
More Important >>
It is important to me
to be an honest and trustworthy colleague
19
SAMPLE CARD:
ACM Ethics Principle: 1.3
Value: 19, *Dependability*

Figure 4. Sample Values Q-Sort – this picture shows how the ACM Code of Ethics principles were mapped to Schwartz’s Values Model (Schwartz 2012 in Winter 2018). The AI Values Q-Sort has been designed in a similar way by using the *Asilomar AI Principles* as statements instead of the ACM Code.

In keeping with this framework, we have designed and developed a selection of tools and techniques for the investigation of values at each level. In this workshop we focus on the use of a Values Q-Sort (Winter 2018) built by mapping Schwartz’s values model (Schwartz 2012) onto the *Asilomar AI Principles*. In previous work (Winter 2018), we used this method to map the ACM Code of Ethics for Software Engineers (Figure 4), and to capture values perceptions of software practitioners in industry and research.

The Q-Sort is an established mixed method that was developed in the 1930s by the psychologist and physicist William Stephenson (Stephenson 1993). It is specifically designed for the systematic study of subjectivity by providing structure to subjective opinions (Watts 2012). The method involves asking participants to sort a series of statements onto a grid according to their level of agreement with each statement.

Q-sorts are usually carried out individually, but for the workshop they will be used in small groups of ~5 people who will jointly decide on the sorting and discuss and note down values alignment and tensions within their own group. Each group will then report their outcomes and discussion to the other groups, while a facilitated discussion will chart alignments and tensions between each group.

Workshop Aim & Objectives

To understand, reflect, articulate and deliberate on the values implications of AI on society. This aim will be delivered through three specific objectives:

1. A base-line shared understanding of AI, including what AI can/cannot do
2. An understanding of the differences and interplay between ethics and human values, and how they may apply to AI/computing in general
3. A practice-based exposure to the articulation and deliberation of human values in AI/computing in general.

Workshop Outline

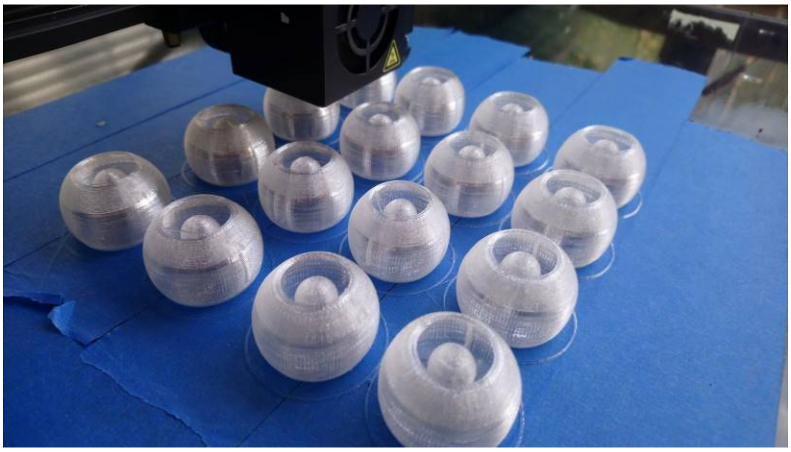


Figure 5. 3DP artifacts may be used as part of the values analysis session

We assume that we will have **90min** to conduct this workshop. This workshop is designed to work with **~30 people** divided in 6 group of 5, but it can also work with larger or smaller groups. We recently used this methods with 75 students attending the European Alpach Forum this summer (EAF2018)

Suggested Final Outcome - the final iteration will involve the co-creation of a 'Prato version' of the *Asilomar AI Principles*.

	Task Description	**Min**
1.	Welcome and intro	5
2.	AI baseline; objective: a high level shared understanding of AI	
-	Bound the scope of the AI to be used in the workshop	10
3.	AI Values Q-sort; objective: group-based articulation and deliberation on AI values priorities (tot:50min)	10
-	Q-sort intro	5
-	Q-sort 1st iteration	20
-	Q-sort 1st group feedback	20
-	Q-sort 2nd iteration; final sort for the Prato version of the AI Principles	
4.	Debrief; 5 min – debrief on tools – how were they designed; wrap up	5
	Tot (min)	**90**

.

Organisers

Maria Angela Ferrario

Lecturer, School of Computing and Communications (SCC), Lancaster University, UK; M. McLuhan Fellow in Digital Sustainability, iSchool, University of Toronto, Canada. Maria Angela has a diverse background including computing (AI/on-line intelligent systems), systems design, social psychology and philosophy. She works at the intersection of software engineering (SE) and human computer interaction (HCI) and is an experienced project manager both within and outside academia. Her research adopts agile and participatory methods to technology development and examines the role of digital technology in society.

Stephen Forshaw

Research Associate, School of Computing and Communications, Lancaster University, UK. Stephen is formerly a Research Associate on the Catalyst project and a Fellow of The Royal Society of Arts. He has a diverse range of professional and academic experience including graphic and product design. Stephen, after a long career in the construction businesses, joined academia and completed his PhD research at the Highwire DTC.

Emily Winter

Research Associate, School of Computing and Communications, Lancaster University, UK. Emily completed her PhD in Sociology at Lancaster University. She was formerly a Research Associate on a Lancaster University/Stanford University research project exploring the values of young people, including in relation to their social media lives. She is interested in values, qualitative research, and technology and society.

Acknowledgements

This work is part of the ‘Values in Computing’ project, funded by the Engineering and Physical Sciences Research Council UK (EPSRC) grant number: EP/R009600/1. Further information: www.valuesincomputing.org .

References

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment–what will keep systems accountable. In 3rd International Workshop on AI, Ethics, and Society.

Asilomar AI Principles. (2017). https://futureoflife.org/ai-principles/ - retrieved on 11th Aug. 2018

Bostrom, N. (2005). Transhumanist values. Journal of philosophical research, 30, 3-14.

Bostrom, N. (2017). Superintelligence. Dunod.

Lynch, S. (2017). https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity - retrieved on 11th Aug. 2018.

Ferrario, M. A., Simm, W., Forshaw, S., Gradinar, A., Smith, M. T., & Smith, I. (2016). Values-first SE: research principles in practice. In Proceedings of the 38th International Conference on Software Engineering Companion (pp. 553-562). ACM.

Ferrario, M. A., Simm, W., Whittle, J., Frauenberger, C., Fitzpatrick, G., & Purgathofer, P. (2017). Values in computing. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 660-667). ACM.

Friedman, B. (1996). Value-sensitive design. interactions, 3(6), 16-23.

Friedman, B., Hendry, D. G., & Borning, A. (2017). A survey of value sensitive design methods. Foundations and Trends® in Human–Computer Interaction, 11(2), 63-125.

Forshaw, S., Cruickshank, L., & Dix, A. (2012). Collaborative communication tools for designing: Physical-cyber environments. In 4th International Workshop on Physicality.

Future of Life Institute (2017) https://futureoflife.org/bai-2017/ - retrieved on 11th Aug. 2018

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In Advances in neural information processing systems (3909-17).

Hanel, P. H., Maio, G. R., Soares, A. K. S., Vione, K. C., Coelho, G. L. D. H., Gouveia, V. V., ... & Manstead, A. S. (2018). Cross-Cultural Differences and Similarities in Human Value Instantiation. Frontiers in psychology, 9, 849

Le Dantec, C. A., Poole, E. S., & Wyche, S. P. (2009). Values as lived experience: evolving value sensitive design in support of value discovery. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1141-1150). ACM.

Maio, G. R., & Olson, J. M. (1998). Values as truisms: Evidence and implications. Journal of Personality and Social Psychology, 74(2), 294.

Maio, G. R. (2010). Mental representations of social values. In Advances in experimental social psychology (Vol. 42, pp. 1-43). Academic Press.

Maio, G.R. (2018) https://theconversation.com/why-society-needs-a-more-scientific-understanding-of-human-values-82537 - retrieved on 11th Aug. 2018.

Newman, P., Ferrario, M. A., Simm, W., Forshaw, S., Friday, A., & Whittle, J. (2015). The role of design thinking and physical prototyping in social software engineering. In IEEE/ACM 37th International Conference on Software Engineering (ICSE).

Riedl, M. O., & Harrison, B. (2016). Using Stories to Teach Human Values to Artificial Agents. In AAAI Workshop: AI, Ethics, and Society.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. Ai Magazine, 36(4), 105-114.

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Pearson Education Limited,

Samar, S. (2018) https://blog.google/products/android/introducing-android-9-pie/ - retrieved on 11th Aug. 2018

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In Advances in experimental social psychology (Vol. 25, pp. 1-65). Academic Press.

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., ... & Dirilen-Gumus, O. (2012). Refining the theory of basic individual values. Journal of personality and social psychology, 103(4), 663.

Stephenson, W. (1993). Introduction to Q-methodology. Operant Subjectivity, 17(1), pp.1-13.

Suchman, L., Follis, K., & Weber, J. (2017). Tracking and Targeting: Sociotechnologies of (In)security. Science, Technology and Human Values, 42 (6), 983

Suchman, L. (2018). https://robotfutures.wordpress.com/2018/04/07/unpriming-the-pump-remystifications-of-ai-at-the-uns-convention-on-certain-conventional-weapons/ - retrieved on 11th Aug. 2018.

Ting, D. S., Liu, Y., Burlina, P., Xu, X., Bressler, N. M., & Wong, T. Y. (2018). AI for medical imaging goes deep. Nature medicine, 24(5), 539.

Van den Hoven, J., Lokhorst, G. J., & Van de Poel, I. (2012). Engineering and the problem of moral overload. Science and engineering ethics, 18(1), 143-155.

ViC Team (2018) Values Tensions in Academia: an Exploration Within the HCI Community ACM interaction magazine June 2018 http://bit.ly/2KlV4l2 - on 11th Aug. 2018.

Watts, S. & Stenner, P. (2012). Doing Q methodological research: Theory, method & interpretation. Sage.

Winter, E., Forshaw, S., & Ferrario, M.A. (2018). Measuring Human Values in Software Engineering. In Proceeding of the 12th International Symposium on Empirical Software Engineering and Measurement. Oct 11-12, 2018 - Oulu, Finland (2018)