Calvo RA, Peters D., Vold V, Ryan, RM (to appear) "Supporting human autonomy in AI systems: A framework for ethical enquiry" in Burr, C. & Floridi, L. (Eds.) Ethics of Digital Well-Being: A Multidisciplinary Approach. Springer Open.

Supporting human autonomy in AI systems: A framework for ethical enquiry

Rafael A Calvo[1,3], Dorian Peters[2, 3], Karina Vold[3, 4], Richard M. Ryan[5]

[1]Dyson School of Design Engineering, Imperial College London, UK. r.calvo@imperial.ac.uk
[2]Design Lab, University of Sydney NSW, Australia
[3]Leverhulme Centre for the Future of Intelligence
[4] Alan Turing Institute, UK
[5] Institute for Positive Psychology and Education, Australian Catholic University, North Sydney, NSW, Australia

Abstract

Autonomy has been central to moral and political philosophy for millenia, and has been positioned as a critical aspect of both justice and wellbeing. Research in psychology supports this position, providing empirical evidence that autonomy is critical to motivation, personal growth and psychological wellness. Responsible AI will require an understanding of, and ability to effectively design for, human autonomy (rather than just machine autonomy) if it is to genuinely benefit humanity. Yet the effects on human autonomy of digital experiences are neither straightforward nor consistent, and are complicated by commercial interests and tensions around compulsive overuse. This multi-layered reality requires an analysis that is itself multidimensional and that takes into account human experience at various levels of resolution. We borrow from HCI and psychological research to apply a model ("METUX") that identifies six distinct spheres of technology experience. We demonstrate the value of the model for understanding human autonomy in a technology ethics context at multiple levels by applying it to the real-world case study of an AI-enhanced video recommender system. In the process we argue for the following three claims: 1) There are autonomy-related consequences to algorithms representing the interests of third parties, and they are not impartial and rational extensions of the self, as is often perceived; 2) Designing for autonomy is an ethical imperative critical to the future design of responsible AI; and 3) Autonomy-support must be analysed from at least six spheres of experience in order to approriately capture contradictory and downstream effects.

Keywords: human autonomy, artificial intelligence, targeting, recommender systems, self-determination theory

1. Introduction

Digital technologies now mediate most human experience from health and education, to personal relations and politics. ‘Mediation’ here refers, not only to facilitation, but also to the ways technologies shape our relations to the environment, including the ways we perceive and behave in different situations. This sense of mediation goes beyond the concept of a technology as a channel of information. It acknowledges that, by changing our understanding of the world and our behaviour, technology affects core features of our humanity. Verbeek (2011), among others, has argued that acknowledging technological mediation is important to understanding the moral dimension of technology, as well as implications for design ethics.

In this paper we focus on human autonomy in relation to technology design ethics. We rely on the definition of autonomy put forward in *self-determination theory* (SDT; Ryan & Deci, 2017) a current psychological theory of motivational and wellbeing psychology. SDT’s approach to autonomy is consistent with both analytic (e.g., Frankfurt, 1971; Friedman, 2003) and phenomenological perspectives (e.g., Pfander, 1967; Ricoeur, 1966) in viewing autonomy as a sense of willingness and volition in acting (Ryan & Deci, 2017). Common in these definitions is viewing autonomous actions as those that are or would be “endorsed by the self”. Critically, according to this definition, autonomy involves acting in accordance with one’s goals and values, which is distinct from the use of autonomy as simply a synonym for either independence or being in control (Soenens et al., 2007). According to SDT one can be autonomously (i.e. willingly) dependent or independent, or one can be forced into these relations. For instance a person can be autonomously collectivisitic, and endorse rules that put group over self (Chirkov et al., 2003). This distinction is significant for our discussion given that, vis-a-vis technologies, individuals may or may not endorse giving over, or alternatively, being forced to retain, control over information or services being exchanged (Peters, Calvo, Ryan; 2018).

The psychological evidence aligned to this conception of autonomy is considerable. From workplaces, to classrooms, to health clinics, to sport fields (Ryan & Deci, 2017), participants who experience more autonomy with respect to their actions have shown more persistence, better performance and greater psychological wellbeing. Evidence for the importance of autonomy-support to human wellbeing, and to positive outcomes more generally, has more recently led to concern about autonomy within technology design (Peters, Calvo & Ryan, 2018). However, the

identification of design strategies for supporting human autonomy poses at least two significant challenges. The first regards breadth: Design for autonomy covers very broad territory given that technologies now mediate experiences in every aspect of our lives and at different stages of human development, including education, workplace, health, relationships and more. The second challenge is that such design practices raise significant ethical questions which can challenge the core of how autonomy has been conceived across multiple disciplines. For example, most technologies are designed to influence (i.e. support or hinder) human behaviours and decision making. As Verbeek (2011) has put it, “Technological artifacts are not neutral intermediaries but actively co-shape people’s being in the world: their perceptions and actions, experience and existence...When technologies co-shape human actions, they give material answers to the ethical question of how to act.” Therefore, intentionally or not, technology design has an impact on human autonomy, and as such, on human opportunities for wellbeing.

This paper elaborates on the nuances of the experience of autonomy within technology environments using a model called METUX (“Motivation, Engagement and Thriving in User Experience”; Peters, Calvo & Ryan, 2018). The model has been described as “the most comprehensive framework for evaluating digital well-being to date” (Burr, Taddeo & Floridi, 2019), and is based on self-determination theory, a body of psychological research that has strongly influenced autonomy-support in fields such as education, parenting, workplaces and health care (Ryan & Deci, 2017). SDT holds that human wellbeing is dependent on the satisfaction of basic psychological needs for autonomy, competence, and relatedness. Herein, we focus exclusively on autonomy owing to its particular relevance in relation to discussions of machine autonomy, and its centrality among principles for ethical AI.

We begin by briefly reviewing some of the predominant conceptions of autonomy within philosophy, giving special attention to notions that stand to inform the design of AI environments. In section 2, we look at how autonomy, and ethics more broadly, have been perceived within the engineering and technology industry. In section 3, we summarise the work in human-computer interaction (HCI) that has bridged technology with the social sciences to improve support for human autonomy within digital systems—sometimes within the larger context of designing for psychological wellbeing. In sections 4 and 5, we provide rationale for the specific value of SDT, as compared to other psychology theories, for understanding AI experience. Then, in section 6 we describe the example of the YouTube video recommender system as a case study for illustrating

various autonomy-related tensions arising from AI, and the value of applying the METUX model for better undertanding the complexities. The model is elaborated in section 6 and applied to the case study in section 7. In section 8, we conclude.

2. Philosophical positions on autonomy

Concepts of human autonomy have long played an important role in moral and political philosophy. Despite general agreement that human autonomy is valuable and merits respect, there is less agreement around what autonomy is, and why (and to what extent) it should be valued and respected. We will not attempt to settle these disagreements, but here we will lay out a few conceptual distinctions with the aim of providing clarity around the notion as we employ it.

The term autonomy was originally used by the Ancient Greeks to characterize self-governing city states. They did not explicitly discuss the concept of individual autonomy, which has, in contrast, preoccupied many modern philosophers. John Stuart Mill, in his famous work *On Liberty*, did not use the term autonomy, but nonetheless argued for the concept of “self-determination” broadly as “the capacity to be one's own person, to live one's life according to reasons and motives that are taken as one's own and not the product of manipulative or distorting external forces.” (Christman 2018). The value of this capacity is not limited to any domain—it is a characteristic that can apply to any aspect of an individual’s life, though for Mill, it is perhaps most significant in the moral and political spheres (Christman, 2018). Indeed, he saw self-determination as a basic moral and political value because it is “one of the central elements of well-being” (Mill 1859/1975, ch. 3). For Mill, then, individual autonomy is a psychological ideal, and represents a constitutive element of one’s well-being. Furthermore, for Mill this ideal has a normative aspect, which grounds certain duties on others. Individuals have a right to self-determine, and so others have an obligation not to unduly interfere with others’ decisions or ability to live in accordance with their own reasons and motives.

Of course, Mill is just one of many philosophers of autonomy. Immanuel Kant, for example, was occupied with an a priori concept of rational autonomy that, he argued, is presupposed by both morality and all of our practical thought. Hill (2013) highlights that in Kant’s view, certain conditions should be met for a decision or action to be considered autonomous. First, the agent has to have certain relevant internal cognitive capacities that are necessary for self-governance, but that are widely thought to be lacking in most animals, children, and some mentally disabled adults.

Second, the individual has to be free from certain external constraints. Like Mill, Kant also recognized that our capacities for rational autonomy can be illegitimately restricted by external forces in many ways, including "by physical force, coercive threats, deception, manipulation, and oppressive ideologies" (Hill, 2013), and that a legal system is needed to "hinder hindrances to freedom" (Kant, *RL*6:230–33; quoted in Hill, 2013). The notion of manipulation and deception as a hindrance to autonomy is particularly relevant within certain technological environments and we will touch on this later within our example.

If, for the sake of this discussion, we accept autonomy as willingness and self-endorsement of one's behaviors, then it's useful to highlight the opposite, *heteronomy*, which concerns instances when one acts out of internal or external pressures that are experienced as controlling (Ryan and Deci 2017). Feeling controlled can be quite direct, as when a technology "makes" someone do something that she does not value (e.g., an online service that forces the user to click through unwanted pages illustrates a minor infringement on autonomy). But it is not only external factors that can be coercive, there are also internally controlling or heteronomous pressures (Ryan, 1982) that can reflect a hindrance to autonomy. For example, technology users can develop a compulsion that leads to overuse, as widely seen with video games and social media (e.g., Przybylski, et al., 2009). Many use the term "addiction," in describing overuse, to convey a coercive quality. Popularly, the concept of FOMO (fear of missing out) describes one such type of technology-induced compulsion to constantly check one's social media. Przybylski, Murayama, DeHaan and Gladwell (2013) found that FOMO was higher in people who heavily used social media, and was also associated with lower basic need satisfaction, including lower feelings of autonomy, and lower mood.

Such examples suggest that even though a user might appear to be opting into a technology willingly, the experience may nonetheless feel controlling. Self-reports that "I can't help it" or "I use it more than I'd like to" reflect behaviour that is not fully autonomous (Ryan & Deci, 2017). In fact, there are now many technologies available which are dedicated solely to helping people regain self-control over their use of other technologies (Winkelman, 2018).

Taking these points together, we can outline a series of characteristics for a conceptualisation of autonomy useful for AI and technology contexts. For this working definition, we can conclude that human autonomy within technology systems requires:

- A feeling of willingingness, volition and endorsement.

- The lack of pressure, compulsion or feeling controlled.
- The lack of deception or deliberate misinformation.

Although this is, of course, not a complete or sufficient conceptualisation for operationalising human autonomy within AI systems, it forms a helpful foundation that provides a basis for addressing a large number of the key tensions that arise within these contexts, which will be demonstrated within our case study in the second half of this chapter. However, first we will turn to perceptions and manifestions of autonomy within computer science, engineering, and human-computer interaction.

3. Notions of autonomy within technology fields

Although we have highlighted that *human* autonomy has long been important to philosophy and the social sciences, engineering and computer science have tended to focus on *machine* autonomy. For example, as of 2019, a search for the word “autonomy” in the Digital Library of the Association for Computing Machinery (ACM) reveals that of the top 100 most cited papers, 90% are on machine autonomy. However, human autonomy has begun to assert itself within the technology industry of late, due to a growing public concern over the impacts of AI on human wellbeing and society. In response, philosophers and technology leaders have gathered and come to concensus over the need to respect and support human autonomy within the design of AI systems (Floridi et al. 2018). New sets of AI principles codify autonomy-support, mirroring a similar refocus on autonomy within health (Beauchamp & Childress, 2013).

The Institute of Electrical and Electronics Engineers (IEEE), the world’s largest professional engineering organisation, states that its mission is to “foster technological innovation and excellence for the benefit of humanity” (IEEE 2019). This benefit has traditionally been interpreted as maximizing productivity and efficiency (i.e. the rate of output per unit of input), an approach that has fuelled decades of work on automation and computer agency within the industry. Automation is a design strategy aimed at maximising productivity by avoiding the need for human intervention. As such, the vast majority of research in engineering has focused on the design of autonomous systems, particularly robots and vehicles (e.g., Baldassarre et al., 2014).

Within engineering practice, there has traditionally been little questioning of productivity, efficiency, and automation as primary strategies for benefiting humanity. Ethics within engineering education has focused on ensuring safe and properly functioning technologies. While it could be

argued that productivity is a poor proxy for human benefit, it might also be argued that, at a basic level, by creating products to satisfy human needs, engineers have taken humans as ends-in-themselves and therefore, essentially acted ethically (in Kantian terms). Yet this would be true only under conditions where the needs satisfied are ones both endorsed and valued by users. In fact, many new business models focus on users data and attention as the basis for monetisation, turning this traditional value structure on its head and make humans merely a “means-to-an-end”. For example, on massively popular platforms like YouTube, Facebook and Instagram, what is being harvested and sold is user attention, which is valuable to marketers of other products. In this new economic model of attention trading, engineers create technologies that collect user data and attention as input, and hours of engagement and user profiling as output to be sold to advertisers. Within these systems, the human is an essential ‘material’ or means to an end.

Aside from some of the broad ethical issues relating to this business model, implications for human autonomy can specifically arise from a disalignment between commercial interests and user interests. Where marketers are the “real” customers, serving user best interest is only important to the extent that doing so is necessary for serving the interests of marketers. Therefore, if there are ways to increase engagement that are manipulative or deceptive to the user, but effective, then these methods are valuable to business (and to the machine learning algorithms programmed to ‘value’ these things and optimise for them).

In addition, when users choose to adopt a technology, but under conditions in which the use of their behavior, personal information, or resources is not disclosed, the user’s autonomy is compromised. This is especially true where the information would potentially alter their choices. Not surprisingly, human autonomy has suffered in a number of ways within this new business model, including through increased exposure to misinformation, emotional manipulation and exploitation. We touch on some of these in more detail in our case study later).

Concerns about this new economy, sometimes referred to as “surveillance capitalism”, have grown steadily (Zuboff, 2019; Wu, 2017). In response, engineers and regulators have begun attempting to devise ethical boundaries for this space. For example, in 2017 the IEEE began the development of a charter of ethical guidelines for the design of autonomous systems that places human autonomy and wellbeing (rather than productivity) at the centre (Chatila et al., 2017). In fact, a growing number of employees and industry leaders, many responsible for contributing to the most successful of the

attention market platforms, are beginning to openly acknowledge the intrinsic problems with these systems and push for more “responsible” and “humane” technologies that better benefit humanity (e.g. humanetech.com; doteveryone.org.uk).

Thus, at the cusp of the third decade of the 21st century, the technology industry finds itself in a kind of ethical crisis with myriad practical implications. Many who benefit from the attention market continue to defend its current strategies, while others are increasingly expressing self-doubt (e.g. Schwab, 2017; Lewis 2019), signing ethical oaths (e.g. the Copenhagen Letter, see Techfestival, 2017), joining ethics committees (see doteveryone.org for a list of charters, oaths and committees), and challenging the status quo within their own organisations (e.g. Rubin, 2018). Others, having identified business models as core to the problem, are experimenting with alternative models, such as subscription services (which generally do not rely on ad revenue), social enterprises, and “B corporations” designed to “balance purpose and profit” (see: http://bcorporation.net/).

4. Designing for autonomy in HCI

A handful of researchers in human-computer interaction have been working on supporting human autonomy through design since at least the 1990s. For example, Friedman (1996) described three key design factors for a user interface that impact autonomy, including system capability, system complexity, misrepresentation, and fluidity. In the last five years, a number of researchers have developed new design methods for supporting autonomy which go beyond the immediate effects of a user interface and extend to autonomy as a life-wide experience. These methods have often approached autonomy through the larger contexts of psychological wellbeing (Peters, Calvo & Ryan, 2018; Gaggioli et al. 2017, Calvo & Peters, 2014; Desmet & Pohlmeyer, 2013; Hassenzahl, 2010) and human values (Friedman & Hendry, 2019, Flanagan & Nissenbaum, 2014) and often build on psychological theories, such as theories of positive psychology (Seligman, 2018), hedonic psychology (Kahneman, Diener & Schwartz, 1999), or motivation (Hekler et al., 2013).

These approaches have generally been based on the idea of translating psychology research into design practice. However, empirical evidence for the effectiveness of these translational models, and the extent to which they impact the quality of design outcomes, is still emerging. Among the psychological theories translated into the design context, SDT has perhaps been the most systematically applied. The likely reasons for this are outlined below.

5. SDT as a basis for autonomy-supportive design

SDT has gathered the largest body of empirical evidence in psychology with respect to issues of autonomy, psychological needs, and wellbeing. In its broadest strokes, SDT identifies a small set of basic psychological needs deemed essential to people's self-motivation and psychological wellbeing. It has also shown how environments that neglect or frustrate these needs are associated with ill-being and distress (Ryan & Deci, 2000; 2017). These basic needs are:

- **Autonomy** (feeling willingness and volition in action),
- **Competence** (feeling able and effective),
- **Relatedness** (feeling connected and involved with others).

Although in this article we focus on the individual’s need for autonomy, we note that aiming to support all three is important for human wellbeing, and therefore, essential criteria for the ethical design of technology. Indeed, innate concerns over our basic psychological needs are reflected in modern anxieties over AI systems. Take, for example, the fears that AI will take over our jobs and skills (threatening our *competence*), take over the world, (threatening our *autonomy*) or replace human-to-human connection (threatening our *relatedness*). Ensuring support for basic psychological needs constitutes one critical component of any ethical technology solution.

In addition to its strong evidence base, there are also a number of qualities of self-determination theory that make it uniquely applicable within the technology context. Firstly, as a tool for applied psychology, SDT is sufficiently actionable to facilitate application to technology and design. However it is not so specific that it loses meaning across cultures or contexts. Up to this point, research on psychological needs across various countries, cultures, and human developmental stages provides significant evidence that autonomy, competence and relateness are essential to healthy functioning universally, even if they are met in different ways and/or valued differentially within different contexts (e.g., Yu, Levesque-Bristol & Maeda, 2018).

Second, SDT literature describes, and provides empirical evidence for, a spectrum of human motivation which runs along a continuum from lesser to greater autonomy (Ryan & Connell, 1989; Howard, Gangne & Breuau, 2018; Litalien et al., 2017). This motivation continnum has, for example, been used to explain varying levels of technology adoption and engagement as well as the powerful pull of video games (Ryan, Rigby & Przybylski, 2005; Rigby & Ryan, 2011).

An additional pragmatic point is that SDT provides a large number of validated instruments for measuring autonomy (as well as wellbeing and motivation). These can be used to directly quantitatively compare technologies or designs with regard to an array of attributes and impacts. Related to this point, perhaps the most important advantage of SDT for integrating wellbeing psychology into the technology context, is its unique applicability to almost any resolution of phenomenological experience. That is to say, its instruments and constructs are as useful at measuring autonomy at the detailed level of user interface controls as they are to measuring the experience of autonomy in someone’s life overall. In contrast, most other theories of wellbeing are applicable only at higher levels. For example, Quality of Life measures used in Wellbeing Economics focus on the life level (Costanza et.al, 2007). Moreover, SDT’s measures can be used to measure the psychological impacts of any technology, regardless of its purpose and whether it is used only occasionally or everyday.

For example, Kerner and Goodyear (2017) used SDT measures to investigate the psychological impact of wearable fitness trackers over eight weeks of use. Results showed significant reductions in need satisfaction and autonomous motivation over that time. Qualitative evidence from focus groups suggested the wearables catalyzed short-term increases in motivation through feelings of competition, guilt, and internal pressure, suggesting some ways in which lifestyle technologies can have hidden negative consequences in relation to autonomy. Furthermore, SDT measures have been widely applied to compare various video game designs, showing how design approaches can differentially impact autonomy, and thereby influence sustained engagement and enjoyment (e.g., Ryan, Rigby & Przybylski, 2005; Peng, et al., 2012).

As helpful as SDT promises to be for technology design research, it has not, until recently, provided a framework for differentiating experiences of autonomy with respect to the various layers of human technology interactions. This gap has only became salient as the theory has been applied in technology applications where a large range of different resolutions must be considered and where these can present contradictory effects on psychological needs. For example, “autonomy-support”, with respect to technology, might refer to customisable settings that provide greater choice in use of the software. Alternatively, it might refer to the way a self-driving car affords greater autonomy in the daily life of someone who is physically disabled. While both describe experiences of increased autonomy, and autonomy-supportive design, they are qualitatively very different and only the latter

is likely to cause measurable impact at a life level. Moreover, a game may increase psychological need satisfaction within the context of gameplay (providing strong experiences of autonomy and competence during play) but hinder these same needs at a life level (if overuse feels compulsive and crowds out time for taking care of work, family and other things of greater import).

Therefore, it is clear that greater precision is required in order to effectively identify and communicate conceptions of autonomy at different resolutions within technology experience. Calvo, Peters, Johnson and Rogers (2014) first highlighted this need and presented a framework distinguishing four “spheres of autonomy”. Peters, Calvo and Ryan (2018) expanded on this work substantially, developing as part of a larger framework, a six-sphere model of technology experience that identifies six distinct levels at which all three psychological needs can be impacted. It is this model that we believe can be usefully applied to our understanding of ethical conceptions of autonomy within technology experiences, and we will describe it in greater detail in section 6. However, it may first be helpful to turn to a case study to provde greater context. Specifically, we provide a brief analysis of the YouTube video recommender system and its implications for human autonomy.

6. Autonomy in context: The example of the YouTube recommender system

Different accounts of autonomy have significantly different practical implications within technology experience. For example, when discussing freedom of speech on the Internet, autonomy is appealed to by both those arguing for the right to free speech (even when it is hateful) and those defending the right to be free *from* hate speech (Mackenzie & Stoljar 2000). The designers of the systems that mediate today’s speech must make values-based decisions that affect this balance, and that impact how individuals experience communication with others.

In the case of YouTube, for example, the action of uploading or ‘liking’ a discriminatory video occurs within the context of a system of recommendations that either supports or suppresses the likelihood of such videos being seen. For example, a user who “likes” one video which contains slightly racially bigoted content, is then likely to get shown more of them, many of which may be more explicitly discriminatory, since the algorithm is influenced by the engagement advantages of extreme and emotionally-charged headlines (i.e. clickbait). Shortly, this user’s YouTube experience may be dominated by videos aligned only to a particular extreme view. This experience leaves the

user within a social “reality” in which “everyone” seems to support what, in truth, may be a very marginal view. “Evidence” is given, not only by the videos that constitute this environment, but also by the thousands of likes associated with each, since the videos have previously been shown primarily to users more likely to “like” them, thanks to the recommendation system.

The ideological isolation caused by this “filter bubble” doesn’t even require the user to enter search terms because recommendations are “pushed” unsolicited into the visual field beside other videos, and may even autoplay. This scenario shows how social influence can be constructed by a system that is deliberately designed to reflect a biased sample. For an unwitting user, this biased representation of the zeitgeist creates reinforcement feedback.

Furthermore, consider the consequences of how frictionless uploading a racially charged video is within systems that create a social environment in which such content would mostly receive positive comments. While in a non-digitally mediated life, a person might not consider producing or engaging with such content because of negative social reactions, in the algorithmically shaped online world the same behaviour is encouraged and perceived as a norm. In other words, before our content was being filtered by an AI system, one had to consider the potential diversity of ‘listeners’. Few would stand on a busy sidewalk in a diverse metropolitan area handing out racially charged flyers. But on YouTube, one can experiment with extreme views, and put out hateful content with some guarantee that it will be shown to an audience that is more likely to receive it well.

The example of YouTube recommender and “like” systems lends strong evidence for the notion of technological mediation (Verbeek, 2011) and the “hermeneutic relations” (Ihde, 1990) through which human interpretation of the world is shaped. The AI-driven recommendation system shapes, not only how we perceive our social situation and our understanding of the world, but also our behaviour. This presents an interesting challenge for autonomy support. Unaware of the bias, a user is likely to feel highly autonomous during the interaction. However, the misinformation (or misrepresentation) potentially represents a violation of autonomy according to most of the philosophical views discussed earlier, as awareness of potentially conflicting alternative information would likely become more phenomenologically salient if the user were informed of the manipulation. Understanding autonomy as reflective endorsement (e.g., Frankfurt, 1971), technologies that obscure relevant considerations compromise autonomy.

This is akin to similar problems within human-human relations, and the definition of ‘misleading’ (as opposed to erroneous) is sometimes controversial since it is often based on intentions which can be difficult to prove. For example, it may benefit technology makers to deliberately obscure information about how data is used (hiding it within inscrutable terms and conditions). For instance, some developers obscure the uses they may make of location data (Gleuck, 2019). In our case study, it’s unlikely YouTube developers deliberately intend to fuel radicalisation. However, they might choose to overlook the effect if the technical approach is sufficiently effective by other measures. While human intention can be difficult to prove, an algorithm’s “intention” is far more straightforward. It must be mathematically defined based on an explicit goal, for example, “optimise user engagement.” This allows for ethical enquiry into the potential consequences of these goals and algorithmic drivers. If we know the algorithm “intends” to do whatever will most effectively increase user engagement and it does so by narrowing the diversity of content shown, what might some of the implications be on human autonomy?

In one sense, YouTube’s system can be thought of as empowering user autonomy—for both producers and consumers of content. It empowers producers to post the content they want to post, while at the same time it is less likely that someone who would be offended will be shown it (freedom to create hate speech is respected while freedom to be free from hate speech is also supported). Indeed, at one level the ‘dilemma’ of hate speech has been resolved (in essence, by creating different worlds and allowing each user to exist in the one they prefer).

But these virtual worlds are illusory and ephemeral and their effects can carry into the real world. We believe a new dilemma arises that can only be clearly seen when viewed across distinct spheres of technology experience. For instance, this optimistic analysis of YouTube’s design as a solution to freedom of speech tensions relies, not only on ignoring the extent to which recommender systems shape the free speech that is viewed, but also on an entirely individualistic and exclusively low-resolution analysis of autonomy--one that excludes the broader social reality of the individual. In this non-relational account, the individual can be considered “autonomous” as long as options are offered and not imposed by the system. However, the system must inevitably “impose” some content to the extent that it can’t show all available videos and must choose on behalf of the user what options they will have. When the number of options is infinite the choice architecture may be driven by social variables. Not taking into account broader social impacts of the technology’s silent restructuring of reality also has consequences.

One example of the consequences of ignoring the socially-situated reality of technologies can be found in the work of Morley and Floridi (2019a, 2019b) who explored the narratives of empowerment often used in health policy. They consider how digital health technologies (DHTs) act as sociocultural products and therefore cannot be considered as separate from social norms or the values they have on others. In this context health technologies designed to “empower” (i.e. support human autonomy) create scenarios of control through which potentially shaming or ‘victim blaming’ messaging fosters introjected motivation, whereby self-worth is contingent on performing the prescribed behaviors (see also Burr & Morley 2019). We argue that a new conceptual lens is needed to make sense of scenarios like these—a lens that considers the different levels at which personal autonomy can be impacted. While any perspective on these questions is likely to be incomplete, we believe that at least acknowledging the various interdependent layers of impact is an important start.

7. Applying the “METUX” model to the analysis of autonomy support within digital experience.

As mentioned previously, self-determination theory posits that all human beings have certain basic psychological needs including a need for competence, relatedness, and, most germane to our discussion, autonomy. Significant evidence for this theory of basic psychological needs (BPNs), has accrued over the past four decades and includes research and practical application in education, sport, health, workplace and many other domains (see Ryan & Deci, 2017; Vansteenkiste, Ryan & Soenens, 2019 for extensive reviews).

Recent efforts applying SDT to technology have revealed the need for an additional framework of analysis in order to more accurately understand BPNs within the technology context. In response, Peters, Calvo & Ryan (2018) developed a model of “Motivation, Engagement and Thriving in User Experience” (METUX). Figure 1 provides a visual representation of the model.

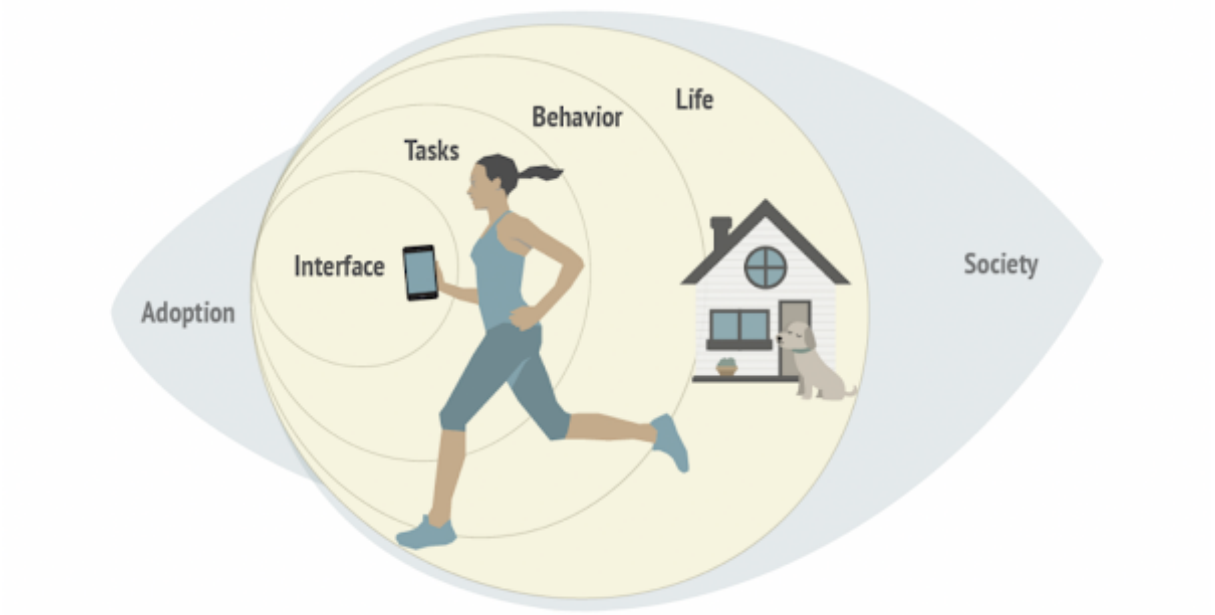

Figure 1: Spheres of Technology Experience, a component of the METUX model.

The METUX model, among other things, introduces six separable “Spheres of Technology Experience” in which a technology can have an impact on our basic psychological needs. Broadly, the first sphere, **Adoption**, refers to the experience of a technology prior to use, and the forces leading a person to use it. For example marketing strategies can tap into internal self-esteem pressures to induce people to buy, or they can take an informational and transparent approach to encourage choice. Adoption can be a function of external and social pressures, or something more volitional.

Once someone begins using a technology, they enter the next four spheres of the “user experience”. At the lowest level of granularity, the **Interface** sphere involves a user’s experience interacting with the software itself, including the use of navigation, buttons and controls. At this level, a technology supports psychological needs largely by supporting competence (via ease-of-use) and autonomy (via task/goal support and meaningful options and controls).

The next sphere, **Task** refers to discrete activities facilitated by the technology, for example “tracking steps” in the case of a fitness app or “adding an event” as part of using calendar software. Separate to the effect of the interface, these tasks can each be accompanied by more or less need satisfaction. Some tasks for example, may feel unnecessary, irrelevant or even forced on users, whereas others are understood as useful, and thus done with willingness.

Combinations of tasks generally contribute to an overall behaviour, and the **Behaviour** sphere encompasses the overarching goal-driven activity enabled, or enhanced, by the technology. For example, the task “step-counting” may contribute to the overall behaviour: “exercise”. Regardless of how need-supportive a technology is at the interface and task levels, a behaviour such as exercising might be more or less a self-endorsed goal and domain of activity.

The final sphere within the user’s direct experience is **Life**, which captures the extent to which a technology influences the fulfillment of psychological needs, such as autonomy, within life overall, thus potentially effecting the extent to which one is “thriving”. For example, even though a person may autonomously adopt an activity “tracker,” and feel comfortable at the interface and task levels, the use of the tracker may still compromise one’s overall sense of autonomy and wellness at the life level, as suggested by the research we reviwed by Kerner and Goodyear (2017).

In sum, a user may feel autonomous when navigating the interface of a fitness app, but not with respect to step counting (e.g. “I can’t possibly do 10,000 steps every day”). Or, they may find step counting increases their sense of autonomy but not their experience of autonomy with regard to exercise overall. Finally, a technology may fulfil psychological needs at the levels of interface, task and behaviour but not have a measurable impact on one’s life. The ways in which the spheres framework allows designers to identify and target need satisfaction at all relevant levels makes them helpful to design. The existence of measures for need satisfaction that can be applied at most of these spheres, also makes them actionable.

Finally, expanding beyond the user experience, we come to **Society** which involves impact on need satisfaction in relation to all members of a society, incuding non-users of a technology (and non-humans). For example, a person might enjoy their new smartphone, and endorse its adoption, but component parts made of gold are manufactured through abusive labour practices. More broadly the volitional use of smartphones may change the overall patterns of interaction between humans, in ways for better and worse or have a collective impact on child development. More detailed explanations for each of these spheres is given in Peters, Calvo and Ryan (2018).

It is important to note that the boundaries between spheres are conceptual and examples of overlap and interrelation naturally exist. The point is not to overemphasize the boundaries but to provide a

way of organising thinking and evaluation in a way that can address the layered, and potentially contradictory, parallel effects of technology designs (e.g., when a technology supports psychological needs at one level while undermining them at another).

8. Returning to the Case Example: Applying the METUX Spheres to YouTube systems.

In the previous section we described the spheres in relation to the satisfaction of psychological needs. Coming back to our YouTube case study, we can begin to apply the lens of the METUX spheres to an exploration of ethical issues to do with autonomy in relation to this technology.

8.1 Adoption

Beginning with Adoption, the initial autonomy-related issue that arises is the extent to which someone’s adoption of a technology is autonomous (rather than controlled). When somone starts using YouTube for the first time, SDT predicts that the extent to which they do so autonomously (i.e. because they wanted to versus because they feel pressured to do so) will have an impact on their engagement afterward. People are often compelled to use technologies, for example, for work, school, or in order to be part of a group or community. While technology makers may have little control over this area of impact, they can look at ways to communicate the benefits of the technology (i.e. through marketing) to increase endorsement. An ethical enquiry might explore questions like: “Do people feel pressured to adopt the platform and if so, what are the sources of that pressure?” “To what extent is information available about the technology’s benefits and risks transparent or misleading?” And, “Is the platform equally available to all people who might benefit from it, or are their exclusionary factors that may be a concern?” (e.g., to do with cost, accessibility, region, etc.).

8.2 Interface

Once someone becomes a user of the platform, we turn to the Interface sphere. Within our example, YouTube’s autoplay feature is an interface design element that can cause autonomy frustration as it makes decisions automatically for the user about what they will watch and when, without confirming endorsement. Autoplay can be turned off, but the feature is opt out rather than opt in. This clearly benefits media providers by increasing hours of user engagement, but the extent to which it benefits users is more questionable and will likely depend on the individual. Autoplay is just one example of how even the design of low-level controls can impact human autonomy and

carry ethical implications. Design for autonomy-support in this interface sphere is largely about providing meaningful controls that allow users to manipulate content in ways they endorse. Focusing on ethics at the interface directs our attention to the things over which users are given control and things over which they are not, as well as the limits placed on that control.

8.3 Tasks

Within the Tasks sphere, we encounter the wide range of activities afforded by a system. Specifically, YouTube supports uploading of videos, “liking” content, searching, browsing, and creating channels, as well as tasks effected by the recommender system described previously. One example of ethical enquiry at this level is provided by Burr, Cristianini, & Ladyman (2018) who review the different ways Intelligent Software Agents (ISA), such as recommender systems, interact to achieve their goals. Specifically, they identify four strategies: coercion, deception, trading and nudging provide *task* level examples such as: “recommending a video or news item, suggesting an exercise in a tutoring task, displaying a set of products and prices”. *Coercion* might involve, for example, forcing a user to watch an ad before continuing to a movie. However, even ‘forced’ behaviours may be relatively endorsed by the user (e.g. “I don’t mind watching an ad if it allows the content to be free”) and designers can work to gain this endorsement by providing rationale for the infringement. *Deception* involves the use of misleading text or images to engage the user in a task (e.g., phishing scams) while *trading* ocurrs when the ISA makes inferences about the users’ goals and uses them to offer options that maximise both the users’ and the ISA’s goals. The final form of interaction presented by the authors is *nudging*, whch involves the use of available information or user bias to influence user decision-making (see Arvanitis, Kalliris, & Kaminiotis, 2019).

In workplaces, tasks are particularly important because they are the focus of automation efforts. While the totality of what an employee experiences as her “job” is often hard to automate, tasks are not. In some cases, task automation can benefit a job, but in others it can be enough to eliminate it. For example, Optical Character Recognition might improve experience for an accountant by making their work more efficient and accurate, however it may entirely eliminate the job of a data entry person. The impact of AI on workplaces will likely be through replacing human *tasks*. Technology designers will often focus on tasks, both when the goal is to engage the user as means-to-a-commercial-end, or when automating something that a human used to do.

8.4 Behaviour

In our YouTube case study, tasks like content browsing and “liking” contribute to different behaviours for different users. Broadly, all users “consume media”, and some of them do this for purposes of “entertainment” or “education”. A smaller number of users, “publish media” and they might do this for the purpose of “communication” or “work,” each of which can be thought of as a behaviour. Targeting autonomy at this level draws attention to the needs of content producers to feel autonomous in creating and disseminating their work and designers might ask “What will help support a video producer’s feelings of autonomy?” or “What are their goals and values and how can YouTube’s design support these?” For an ethical enquiry, we might investigate what rights producers retain with respect to their content, what policies and limits are placed on what can be published, as well as the reasons for those limits. We might also scruitinize the ways media is presented or distorted as a result of the unique characteristics of the technology, and what implications this might have on the autonomy of users.

Moreover, the way in which technologies work to grab attention is critical to ethical questions of autonomy, since, if we accept that attention, as William James (1890) described it, is “the essential phenomenon of will” there is little room for autonomous action without it. For example, when a student watches a lecture on Youtube for a class, he is pursuing a goal to learn and fulfil course requirements. When his attention is then drawn by a video recommendation, the original intention (to learn) may be forgotten, and with it, the nature of the behaviour. Behaviour change is often driven by an intention imposed by a technology, and often without awareness of the individual effected, and therefore can be said to affect autonomy.

8.5 Life

In some cases, YouTube may become a significant influence on someone’s life in either a positive or negative way. For example, one user might earn enough to make a living as a “YouTuber” while another may start and maintain a yoga practice because of it. On the other hand, another user may find it difficult to stop overusing YouTube, or a vulnerable teenager may find herself with easy access to pro-anorexia videos.

As we touched on previously, designing to increase the amount of time users spend on a system can fuel overuse, reducing time they have available to engage in other healthy activities (such as connecting with friends, parenting, working, or experiencing nature). This can have consequences on life-level autonomy and other psychological needs. In extreme cases, overengagement has been

viewed as addiction (Kuss & Lopez-Fernandez, 2016), a condition in which autonomy is significantly frustrated.

The possible examples are many but the important point is that circumstances exist in which YouTube will have measurable effects on autonomy at the *Life* level. Ethical enquiry into life-level impact explores influence above and beyond the virtual boundaries of the technology and will rely on research into the human experience of actual use or, for new or prototype technologies, on anticipatory analysis.

8.6 Society

Finally, should some of these life level expriences propogate they could add up to identifiable impact within the *society* sphere. Here again, a combination of sociological research on patterns of use and/or anticipatory processes involving multiple stakeholders will be necessary for identifying critical ethical issues which stand to reverberate across society.

A useful parallel might be drawn with respect to ‘sustainable’ and ‘circular’ design. Just as we need to design in ways that preserve the natural environment for our survival, digital technologies, like YouTube, need to be designed in ways that minimise negative impact on individuals and societies to preserve a ‘sustainable’ *social* environment. For example, the extent to which recommendation systems might coopt attention, change intention and behaviour, and even construct social norms, could have deleterious effects on social interaction, societal values and politics. Filter bubble dynamics, discussed earlier, may deprive individuals of contact with information that may influence their reflective considerations, leading them to support social movements they otherwise would not endorse. Finally, technologies may drive consumer behaviors which may be satisfying in an immediate sense, but which ultimately impact the health and wellness of many members of society, including those who do not consume, or cannot access, the products pushed by a technology.

Addressing societal autonomy requires a relational conception of autonomy (Mackenzie & Stoljar, 2000) which acknowledges, among other things, the extent to which individual autonomy is socially situated and therefore influenced by willing obligations and interdependence with others (e.g., caring between parents and children, the collective goals of a group, a desire for national sovereignty.) When a child’s wellbeing is negatively affected by a technology, it is also the parent’s autonomy that suffers. When fairness is undermined by algorithmic bias, it is a segment of

society whose autonomy may be effected. When democracy is undermined by the generation and targeting of fake news, national autonomy may be threatened.

We argue that, in order for AI products to be considered responsible, and to, therefore, be successful in the longer term, they need to consider their impact within all of the above mentioned spheres—including life and society—both by anticipating potential impact, and then evaluating it regularly once the technology is in use. In table 1 we summarise various types of impact on autonomy arising from the use of YouTube and present these against the METUX spheres of technology experience.

Table 1: Spheres of Technology Experience for YouTube with examples of factors likely to impact autonomy in each.

Sphere of Experience	Support for autonomy	Compromise to autonomy
Adoption *To what extent is technology adoption autonomously motivated?*	Most users adopt YouTube autonomously as it is primarily used for entertainment and self-guided learning, rather than as an obligatory tool for work or communication.	Some users (publishers) may feel pressured to use YouTube over other video platforms (e.g. Vimeo) owing to market dominance.
Interface *To what extent does direct interaction with the technology (i.e., via the user interface) impact autonomy.*	"10 seconds back" and "skip ad" buttons allow users more refined control over content. Controls are also provided for adjusting data input to recommendation systems.	There is no way to skip the beginning of ads (coercive); Videos will autoplay (without user consent) unless the setting is turned off (an opt out).
Tasks *What are the technology specific tasks? How do they impact on autonomy?*	Tasks such as subscribing to channels and 'liking' allow users to customise content. Searching provides access to nearly endless content options.	Deception through clickbait leads to unintended activity; Recommender system results limit options, may distort social norms and may change behaviours online and offline.
Behaviour *How does the technology impact autonomy with respect to the behaviour it supports?*	YouTube contributes to users' ability to engage in a number of behaviours, for example, for educate or entertain themselves. Others are able to share media in order to communicate, work, or engage in a hobby in whole new ways.	Strategies for increasing user engagement increase the risk of overuse or "addiction". Some "educational" content on YouTube may be deliberately or inadvertently misleading. Users may not be aware of how YouTube uses the media they uploaded to it (and what rights they retain).

Life *How does the technology influence the user's experience of autonomy in life overall?*	Greater opportunities for entertainment, education and work flexibility can have an impact on one's overall life.	Instances of radicalization exist. Some videos may promote unhealthy or dangerous behaviours.
Social *To what extent does the technology impact on experiences of autonomy beyond the user and across society?*	People have more potential to communicate, find like others, and organise. Societal trends are formed and shaped.	Due to its reach, YouTube videos can influence public opinion and politics, and rapidly spread sources of disinformation.

9. Discussion: Ethics in the design of AI systems.

In this chapter we have described how the METUX model's "Spheres of Technology Experience" might contribute to clearer thinking, analysis and design in relation to human autonomy within AI systems. We have proposed that the spheres present a useful starting point for applying necessary dimensionality to these discussions. The METUX model also provides instruments that could be used to measure the differential impacts of different design decisions on users at each level. In order to illustrate this point, we described how the model might be applied in the context of YouTube, a familiar AI case study.

In conclusion, if we are to be guided by both philosophers and psychologists with regard to an ethical future for technology, than there is no way forward without an understanding of human autonomy and ways to safeguard it through design. Understanding the phenomenological experience of autonomous behaviour as well as the multifaceted and layered ways in which users of technologies can be controlled or supported in acting autonomously (sometimes in parallel) are essential. Pursuit of this understanding must proceed at both a universal and at context-specific levels as patterns will exist across many technologies, yet each implementation of AI will also have a unique set of contextual issues specific to it. Knowledge in these areas will contribute to informing evidence-based strategies for (more ethical) autonomy-supportive design. In sum, we hope the work presented herein can help contribute to a future in which technologies that leverage machine autonomy do so to better support human autonomy.

Acknowledgements and Reported interests

RAC and DP have received payment for providing training on wellbeing-supportive design to Google. KV was supported by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015-067, and by the Digital Charter Fellowship Programme at the Alan Turing Institute, UK. RMR is a co-founder of Immersyve Inc., a motivation consulting and assessment firm.

References

Arvanitis, A., Kalliris, K., & Kaminiotis, K. (2019). Are defaults supportive of autonomy? An examination of nudges under the lens of Self-Determination Theory. *The Social Science Journal*. https://doi.org/10.1016/j.soscij.2019.08.003

Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M., & 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

Beauchamp, T.L. & Childress, J.F. (2013). Principles of biomedical ethics. 7th ed. New York: Oxford University Press.

Burr, C., & Morley, J. (2019). Empowerment or Engagement? Digital Health Technologies for Mental Healthcare. (May 24, 2019). Available at SSRN: https://ssrn.com/abstract=3393534

Burr, C., Taddeo, M., & Floridi, L. (2019). The Ethics of Digital Well-Being: A Thematic Review. (February 7, 2019). Available at SSRN: https://ssrn.com/abstract=3338441

Burr, C., Cristianini, N., & J. Ladyman. (2018). An Analysis of the Interaction Between Intelligence Software Agents and Human Users. *Minds and Machines*, 28(4): 735-774.

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

Chatila, R., Firth-Butterflied, K., Havens, J. C., & 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

Chirkov, V., Ryan, R. M., Kim, Y., & Kaplan, U. (2003). Differentiating autonomy from individualism and independence: A self-determination theory perspective on internalization of cultural orientations and well-being. *Journal of Personality and Social Psychology*, *84*(1), 97–110.

Christman, J. "Autonomy in Moral and Political Philosophy", The Stanford Encyclopedia of Philosophy (Spring 2018 Edition), Edward N. Zalta (ed.), Available online <https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/>

Christman, J. (Ed) (1989). *The Inner Citadel: Essays on Individual Autonomy*, New York: Oxford University Press.

Costanza, R., Fisher, B., Ali, S., Beer, C., Bond, L., Boumans, R., ... & Gayer, D. E. (2007). Quality of life: An approach integrating opportunities, human needs, and subjective well-being. *Ecological economics*, *61*(2-3), 267-276.

Deady, M., Johnston, D., Milne, D., Glozier, N., Peters, D., Calvo, R., & Harvey, S. (2018). Preliminary Effectiveness of a Smartphone App to Reduce Depressive Symptoms in the Workplace: Feasibility and Acceptability Study. *JMIR MHealth and UHealth*, *6*(12), e11661. https://doi.org/10.2196/11661

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

Diener, E., Inglehart, R., & Tay, L. (2013). Theory and Validity of Life Satisfaction Scales. *Social Indicators Research*. https://doi.org/10.1007/s11205-012-0076-y

Floridi, L. *et al.* (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machine* 28, pp. 689-707. Available online: https://link.springer.com/article/10.1007/s11023-018-9482-5

Frankfurt, H.G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, 68, 5-20.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.

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

Friedman, M. (2003) *Autonomy, gender, politics*. New York: Oxford Univ. Press

Gaggioli, A., Riva, G., Peters, D., & Calvo, R. A. (2017). Chapter 18 - Positive Technology, Computing, and Design: Shaping a Future in Which Technology Promotes Psychological Well-Being. In *Emotions and Affect in Human Factors and Human-Computer Interaction* (pp. 477–502). https://doi.org/https://doi.org/10.1016/B978-0-12-801851-4.00018-5

Gleuck, J. (2019, Oct. 16). How to stop the abuse of location data. *New York Times.Com*

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

Hekler, E. B., Klasnja, P., Froehlich, J. E., and Buman, M. P. (2013). Mind the theoretical gap: interpreting, using, and developing behavioral theory in HCI research. *Proc. CHI* 2013, 3307–3316. doi: 10.1145/2470654. 2466452

Hill, T. (2013). Kantian Autonomy and Contemporary Ideas of Autonomy. In Kant on Moral Autonomy, Ed. Oliver Sensen. Cambridge University Press. pp. 15-31.

IEEE 2019, "Vision and Mission" https://www.ieee.org/about/vision-mission.html (Accessed 21 Oct 2019)

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth* (No. 560). Indiana University Press. Chicago

Institute of Electrical and Electronics Engineers (IEEE). (2019). *Mission and Vision,* IEEE. Retreived on 13 October, 2019. <https://www.ieee.org/about/vision-mission.html>

James, W. (1890). *The Principles of Psychology, Volumes I and II*. 1983 edition. Cambridge, MA: Harvard University Press.

Kahneman, D., Diener, E., & Schwarz, N. (1999). Well-being: The foundations of hedonic psychology. *Health San Francisco*. https://doi.org/10.7758/9781610443258

Kerner, C. & Goodyear V.A. (2017): The Motivational Impact of Wearable Healthy Lifestyle Technologies: A Self-determination Perspective on Fitbits With Adolescents, American Journal of Health Education, DOI: 10.1080/19325037.2017.1343161

Kuss, D. J., & Lopez-Fernandez, O. (2016). Internet addiction and problematic Internet use: A systematic review of clinical research. *World journal of psychiatry*, *6*(1), 143–176. doi:10.5498/wjp.v6.i1.143

Litalien, D., Morin, A. J. S., Gagné, M., Vallerand, R. J., Losier, G. F., & Ryan, R. M. (2017). Evidence of a continuum structure of academic self-determination: A two-study test using a bifactor-ESEM representation of academic motivation. *Contemporary Educational Psychology, 51*, 67-82.

Lewis, Paul (2019). At: https://www.theguardian.com/technology/2017/oct/05/smartphone-addiction-silicon-valley-dystopia. Accessed on: 5/9/2019

Mackenzie, C., & Stoljar, N. (Eds.). (2000). Relational autonomy: Feminist perspectives on autonomy, agency, and the social self. Oxford University Press.

Mill, J.S. (1859/1975). *On Liberty*, David Spitz, ed. New York: Norton.

Morley, J., & Floridi, L. (2019a). The Limits of Empowerment: How to Reframe the Role of mHealth Tools in the Healthcare Ecosystem. Science and engineering ethics, 1-25.

Morley, J., & Floridi, L. (2019b). Enabling digital health companionship is better than empowerment. The Lancet Digital Health.

Owens, J., & Cribb, A. (2013). Beyond choice and individualism: understanding autonomy for public health ethics. Public Health Ethics, 6(3), 262-271.

Peng, W., Lin, J.-H., Pfeiffer, K. A., and Winn, B. (2012). Need satisfaction supportive game features as motivational determinants: an experimental study of a self-determination theory guided exergame. *Media Psychol*. 15, 175–196. doi: 10.1080/15213269.2012.673850

Peters, D., Calvo, R.A., & Ryan, R.M. (2018). “Designing for Motivation, Engagement and Wellbeing in Digital Experience” *Frontiers in Psychology* – Human Media Interaction. Vol 9. pp. 797.

Pfander, A. (1967). Motive and motivation. Munich: Barth, 3rd ed., 1963 (1911); Translation in Phenomenology of Willing and Motication. H. Spiegelberg (ed.) Evanston, IL: Northwestern University Press, 1967.

Przybylski, A. K., Murayama, K., Dehaan, C. R., & Gladwell, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2013.02.014

Przybylski, A. K., Weinstein, N., Ryan, R. M., & Rigby, C.S. (2009). Having to versus wanting to play: Background and consequences of harmonious versus obsessive engagement in video games. *CyberPsychology & Behavior*, *12*(5), 485-492. doi: 10.1089/cpb.2009.0083.

Owens, J. & Cribb, A. (2013). Beyond Choice and Individualism: Understanding Autonomy for Public Health Ethics, *Public Health Ethics*, 6(3): 262–271. https://doi.org/10.1093/phe/pht038

Ricoeur, P. (1966). Freedom and nature: The voluntary and involuntary. (Trans. E. V. Kohák). Evanston, IL: Northwestern University Press.

Rigby, S., & Ryan, R. M. (2011). *Glued to Games: How Video Games Draw us in and Hold us Spellbound*. Santa Barbara, CA: Praeger.

Rubin, B.F. (2018) “Google employees push back against company's Pentagon work”, CNET http://www.cnet.com/news/google-employees-push-back-against-companys-pentagon-work 4/4/18. Accessed at: 6/9/2019

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

Ryan, R.M., Rigby, C.S. & Przybylski, A. Motiv Emot (2006) 30: 344. https://doi.org/10.1007/s11031-006-9051-8

Ryff, C. D. (2018). Well-Being With Soul: Science in Pursuit of Human Potential. *Perspectives on Psychological Science*, *13*(2), 242–248. https://doi.org/10.1177/1745691617699836

Seligman, M. (2018). PERMA and the building blocks of well-being. *Journal of Positive Psychology*. https://doi.org/10.1080/17439760.2018.1437466

Schwab, K. (2017). “Nest Founder: I Wake Up In Cold Sweats Thinking, What Did We Bring To The World?” Fast Company. 7/7/2017. https://www.fastcompany.com/90132364/nest-founder-i-wake-up-in-cold-sweats-thinking-what-did-we-bring-to-the-world. Accessed on 6/9/2019

Soenens, B., Vansteenkiste, M., Lens, W., Luyckx, K., Goossens, L., Beyers, W., & Ryan, R. M. (2007). Conceptualizing parental autonomy support: Promoting independence versus promoting volitional functioning. *Developmental Psychology*, *43*(3), 633-646. doi: 10.1037/0012-1649.43.3.633

Techfestival. (2017). *The Copenhagen Letter*, Techfestival, Copenhagen. Retrieved on 13 October, 2019. <https://copenhagenletter.org>

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

Yu, S., Levesque-Bristol, C., & Maeda, Y. (2018). General need for autonomy and subjective well-being: A Meta-analysis of studies in the US and East Asia. *Journal of Happiness Studies, 19*(6), 1863-1882.

Vansteenkiste, M., Ryan, R.M. & Soenens, B. (2019). Basic Psychological Need Theory: Advancements, Critical Themes, and Future Directions. *Motivation and Emotion*, Advance Online Publication.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Winkelman, S. (2018). “The Best Apps for Limiting Your Screen Time.” *Digital Trends.* January 6, 2018. Accessed 6/9/19 at: https://www.digitaltrends.com/mobile/best-apps-for-limiting-your-screen-time/

Wu, T. (2017) The Attention Merchants: The Epic Scramble to Get Inside our Heads.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books