# Syllabus CT5142 AI and Ethics (provisional – 02/09/2022)

Tuesday 9-11am (IT207)

Lecturer: Dr Heike Felzmann

Contact: [heike.felzmann@universityofgalway.ie](mailto:heike.felzmann@universityofgalway.ie)

This module aims to provide students with a critical understanding of current and emerging ethical concerns in relation to Artificial Intelligence and the digital sphere more generally. Students will be introduced to core concerns in the field and will discuss emerging concerns with a view to their relevance for citizens and professionals in the field of information technology. Preparation of weekly readings is an essential requirement for participation.

## Learning Outcomes

On successful completion of this module the learner will be able to:

LO1 Describe practical concerns in the area of AI and information technology ethics

LO2 Identify ethical concerns arising within the context of AI and information technology ethics

LO3 Critically analyse proposed solutions to those challenges

LO4 Display the ability to search and select relevant professional literature on online databases

LO5 Actively take part in constructive and critical class reflections

LO6 Display a coherent understanding of a defined ethical issue on the basis of independent use of relevant professional sources

LO7 Create effective communication materials on this ethical issue for a defined target audience

## Indicative Content: Overview

Embedded values, value-sensitive design and trustworthy AI

Privacy, consent and dark patterns

Understanding the value(s) of privacy

Datafication, profiling and algorithmic bias

Algorithmic governance and accountability

Surveillance capitalism

Virtual and augmented reality and deepfakes

## Indicative content: Week by week

**Week 1: Embedded values and the trustworthy design of AI**

This week addresses the question whether information technologies embody values, and how technology development might take such values into account. The notion of embedded values will be discussed, and a design approach called “Value-sensitive design” will be introduced, together with some recent attempts targeted at information technology professionals to guide design of information technologies and AI will be presented (Privacy by Design, HLEG AI 2019 Guidance on Trustworthy AI, IEEE Ethically Aligned Design). It will be reflected on whether these approaches will contribute to increased trustworthiness of AI.

Cavoukian, A. (2009). Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada*, <https://www.ipc.on.ca/wp-content/uploads/Resources/pbd-implement-7found-principles.pdf>

Georgieva, I., Lazo, C., Timan, T., & van Veenstra, A. F. (2022). From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. *AI and Ethics*, 1-15.

High-Level Expert Group on AI (2019). *Ethics Guidelines for Trustworthy AI*. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

IEEE (2018) *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (V2)*, <https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf>

Nissenbaum, H. (2001). How computer systems embody values. *Computer*, *34*(3), 120-119.

SIENNA Project (2020). *Ethics by Design and Ethics of Use in AI and Robotics*, <https://sienna-project.eu/digitalAssets/915/c_915554-l_1-k_sienna-ethics-by-design-and-ethics-of-use.pdf>

Umbrello, S., and van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics* 1, 283-96.

**Week 2: Privacy, consent and dark patterns**

This week we will introduce privacy as a concept, looking at privacy through the lens of consent, given that this is how most of us encounter privacy most frequently in the digital sphere. According to the GDPR data processing for many purposes needs to be consent-based, i.e. the user has to explicitly agree to their data being processed. This usually means that users are asked to accept privacy policies in order to access online services – but does that mean they have truly given informed consent? We are going to look at how informed consent is usually understood from an ethical perspective and will compare criteria for informed consent to what happens in the acceptance of privacy notices for digital services. We will look at various voices who are sceptical about the success of the current approach to privacy notices, explore the notion of “dark patterns” and “nudges” and consider different ways of enhancing users’ privacy understanding and privacy practices. We will look at Google’s consent documentation in light of these discussions and will reflect on how they manage the presentation of consent.

Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L. F., Komanduri, S., ... & Wang, Y. (2017). Nudges for privacy and security: Understanding and assisting users’ choices online. *ACM Computing Surveys (CSUR)*, *50*(3), 1-41.

Ben-Shahar, O., Schneider, C. (2010). The Failure of Mandated Disclosure, University of Chicago Law & Economics, Olin Working Paper No. 516, <https://www.law.umich.edu/centersandprograms/lawandeconomics/abstracts/2010/Documents/10-008benshahar.pdf>

Calo, R. (2011). Against notice skepticism in privacy (and elsewhere). *Notre Dame Law Review*, *87*, 1027-1072.

Google Privacy Policy**,** <https://policies.google.com/privacy>

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. In Proceedings of the 2018 CHI conference on human factors in computing systems (pp. 1-14).

Luguri, J., & Strahilevitz, L. J. (2021). Shining a light on dark patterns. *Journal of Legal Analysis*, *13*(1), 43-109.

**Week 3: Understanding the value(s) of privacy**

We will look into some important components of privacy, including Solove’s “Taxonomy of Privacy” and reflect on the so-called “privacy paradox” and whether privacy should be considered obsolete. We will then discuss one influential theory on privacy, Nissenbaum’s “Privacy as Contextual Integrity”, a theory that draws on the idea of different “spheres” in our lives being governed by different norms. Nissenbaum argues that theories of privacy should not be expected to classify information as either private or public, but need to take into account person’s expectations of what contextual norms of information flow apply in a context and how technologies mediate information availability.

Hartzog, W. (2017). The inadequate, invaluable fair information practices. Maryland Law Review, 76(4), 952-982.

Nissenbaum, H. (2019). Contextual integrity up and down the data food chain. *Theoretical inquiries in law*, *20*(1), 221-256.

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, *140*(4), 32-48.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Solove, D. J. (2005). A taxonomy of privacy. *University of Pennsylvania Law Review*, *154*, 477-564.

Solove, D. J. (2007). I've got nothing to hide and other misunderstandings of privacy. *San Diego Law Review*, *44*, 745-772.

Solove, D. J. (2021). The myth of the privacy paradox. *George Washington Law Review*, *89*, 1-41.

Zimmer, M. (2008). Privacy on planet Google: Using the theory of contextual integrity to clarify the privacy threats of Google's quest for the perfect search engine. *Journal of Business & Technology Law*, *3*, 109-126.

**Week 4: Datafication, profiling and algorithmic bias**

This week we are discussing what happens when more information about ourselves becomes digitally available to ourselves and other parties. The notion of “datafication” and the “digital double” will be introduced, and it will be explored what roles and shortcomings digital representations of ourselves or of certain aspects of ourselves may have on our understanding ourselves (e.g. through fitness trackers). We will then explore how digital doubles serve to profile us online, with potentially far-reaching inferences drawn about ourselves by online data collectors, such as Google, Facebook or TikTok. This profiling serves to provide us with “relevant content”, but can also be used for more significant purposes – and what happens if these inferences about us are wrong? We will then explore algorithmic bias as one important challenge in this context, the problem that algorithmic decision-making may be systematically inaccurate for many underrepresented groups, and will discuss the significance of this bias and potential ways of addressing algorithmic bias.

Barocas, S., & Levy, K. (2020). Privacy dependencies. *Washington Law Review*, 95, 555.

Johnson, G. M. (2021). Algorithmic bias: on the implicit biases of social technology. *Synthese*, *198*(10), 9941-9961.

Lanzing, M. (2016). The transparent self. *Ethics and Information Technology*, *18*(1), 9-16.

Lupton, D. (2016). *The quantified self*. John Wiley & Sons.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big data*, *1*(2), 85-99.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review*, 494.

White House Report, Executive Office of the President (2016). Big Data: A report on algorithmic systems, opportunity and civil rights. <https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf>

**Week 5: Algorithmic governance and accountability**

This week we are going to reflect on what happens when our societal interactions become increasingly underpinned by widespread data capture and algorithmic decision-making. We will discuss the notion of algorithmic governance, and the problems arising from the use of automatic processes that assist or replace human beings in the decision-making process, especially those that operate as “black boxes” without being able to explain in detail how they are functioning. As Eubanks argued, these can be particularly problematic when they are used with regard to marginalised communities.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, *20*(3), 973-989.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review*, *8*(4), 1-18.

Lyon, D. (2017). Digital citizenship and surveillance| surveillance culture: Engagement, exposure, and ethics in digital modernity. *International Journal of Communication*, *11*, 19.

**Week 6: Surveillance Capitalism**

After reflecting on different forms of surveillance we will discuss Zuboff’s theory of “Surveillance Capitalism”. Zuboff argues that contemporary society is characterised by transition to a new kind of capitalism that operates through the use of new forms of exercising power. It relies on ever more fine-grained surveillance of our actions especially by the big platforms, such as Google and Facebook, that can be harnessed to influence and manipulate our thoughts, emotions and behaviour in an unprecedented manner.

Darmody, A., & Zwick, D. (2020). Manipulate to empower: Hyper-relevance and the contradictions of marketing in the age of surveillance capitalism. *Big Data & Society*, *7*(1), 2053951720904112.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75-89.

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

**Week 7: Scale, power and monopolies**

This week we are going to reflect on the phenomenon of huge global internet companies that have unprecedented amounts of data at their disposal with unprecedented opportunities to affect societal processes. We are going to explore considerations around specific challenges associated with regulating the power of internet monopolies, drawing on work by Lina Kahn, the current Chair of the FTC.

Alford, R. P. (2022). The bipartisan consensus on big tech. Emory Law Journal, 71(5), 893-932.

Khan, L. (2017). Amazon's antitrust paradox. *Yale Law Journal*, *126*.

Khan, L. M. (2017). Sources of tech platform power. *Georgetown Law Technology Review*, *2*, 325.

Srinivasan, D. (2019). The antitrust case against Facebook: A monopolist's journey towards pervasive surveillance in spite of consumers' preference for privacy. *Berkeley Business Law Journal*, *16*, 39.

**Week 8: Relational artificial agents**

In our daily life we increasingly interact with artificial agents, knowingly or unknowingly. We may be engaging with customer service bots, with Alexa or Siri, with predatory bots on social media, or perhaps with toy robots or service robots. This week we are going to reflect on what it means to engage in interaction with increasingly complex virtual agents, or even robots. On the one hand, many people enjoy interactions with autonomous agents that are designed to serve them, on the other hand, there is a concern whether these relationships are ethically relevant, and perhaps potentially ethically problematic, due to the lack of genuine reciprocity in these artificial relationships. One of the ethical concerns arising here is the societal place given to relational artificial agents and the nature of relationships that people might form with them.

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and information technology*, *14*(1), 27-40.

Sorell, T., & Draper, H. (2014). Robot carers, ethics, and older people. *Ethics and Information Technology*, *16*(3), 183-195.

Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction studies*, *8*(3), 501-517.

Vallor, S. (2011). Carebots and caregivers: Sustaining the ethical ideal of care in the twenty-first century. *Philosophy & Technology*, *24*(3), 251.

**Week 9: AI, creativity, and moral responsibility**

This week we are going to reflect on what issues might arise when AI is used for purposes that human beings consider uniquely human, including for example creativity, moral decision-making or highly sensitive purposes, such as killing or caring for vulnerable persons (as already discussed last week). We are going to reflect on what the easy availability of AI that can create convincing original texts (such as GPT-3) or artworks (such as DALL-E) may mean for our understanding of what were previously considered distinctively human capabilities. We will also consider what responsibilities might come with deploying AI or information technologies with significant impact on human lives, with regard to the question of moral responsibility.

One frequently discussed question is how to understand responsibilities of autonomous agents, e.g. in the case of autonomous cars; we will explore this question of responsibility with regard to its various conceptual components and concerns.

Alaieri, F., & Vellino, A. (2016). Ethical decision making in robots: Autonomy, trust and responsibility. In *International conference on social robotics* (pp. 159-168). Springer, Cham.

Coeckelbergh, M. (2021). Three responses to anthropomorphism in social robotics: Towards a critical, relational, and hermeneutic approach. *International Journal of Social Robotics*, 1-13.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299-309.

Donath, J. (2020). Ethical issues in our relationship with artificial entities (pp. 53-73). Oxford, UK: Oxford University Press.

Fossa, F. (2018). Artificial moral agents: moral mentors or sensible tools?. *Ethics and Information Technology*, *20*(2), 115-126.

Noor, N., Rao Hill, S., & Troshani, I. (2021). Artificial intelligence service agents: role of Parasocial relationship. *Journal of Computer Information Systems*, 1-15.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics*, *24*(4), 1201-1219.

**Week 10: Virtual reality (VR), augmented reality (AR) and deepfakes**

Increasingly, virtual and augmented reality applications become more accessible and integrated into everyday activities. Many people own VR headsets and play VR games. AR applications are increasingly available for professional, personal (shopping aids for furniture, clothes or glasses) or purely recreational uses (e.g. beauty filters). The technological developments underpinning these have also led to the availability of deepfakes, manipulating videos so that they depict real persons seemingly saying things or acting in a particular manner that they never did in reality. This increasing challenge of having representations of reality available and interact with those representation that do not match shared physical reality open up new opportunities for human experience but also raise ethical and political concerns. This week we will be reflecting on potentially ethically beneficial uses of AR and VR as well as various ethical concerns, with particular focus on everyday applications like beauty filters and the potential personal and political risks of not easily being able to distinguish whether a representation of reality is manufactured.

De Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy & Technology*, *34*(4), 1311-1332.

Neely, E. L. (2019). Augmented reality, augmented ethics: who has the right to augment a particular physical space?. *Ethics and Information Technology*, *21*(1), 11-18.

O’Brolcháin, F., Jacquemard, T., Monaghan, D., O’Connor, N., Novitzky, P., & Gordijn, B. (2016). The convergence of virtual reality and social networks: threats to privacy and autonomy. *Science and engineering ethics*, *22*(1), 1-29.

Rueda, J., & Lara, F. (2020). Virtual reality and empathy enhancement: Ethical aspects. *Frontiers in Robotics and AI*, 160.

Ryan-Mosley, T. (2021). Beauty filters are changing the way young girls see themselves, MIT Technology Review, <https://www.technologyreview.com/2021/04/02/1021635/beauty-filters-young-girls-augmented-reality-social-media/>

Slater, M., Gonzalez-Liencres, C., Haggard, P., Vinkers, C., Gregory-Clarke, R., Jelley, S., ... & Silver, J. (2020). The ethics of realism in virtual and augmented reality. *Frontiers in Virtual Reality*, *1*, 1.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11).

Wolfendale, J. (2007). My avatar, my self: Virtual harm and attachment. *Ethics and information technology*, *9*(2), 111-119.

**WK11: Automation, work and human replacement**

AI and robots with their increasing capabilities are employed more and more to take over work that has been previously completed by humans, either completely replacing human labour or supporting human tasks. This has raised concerns with regard to the future of human work, including the question whether significant parts of human labour may become obsolete in the near future, and whether collaboration with AI and robots might enhance or diminish the quality of work for those affected. Another relevant concern is the rise of platform labour which has optimised delivery of services but substantially eroded working conditions for workers.

European Commission: Final report of the High-Level Expert Group on the Impact of the Digital Transformation on EU Labour Markets, <https://digital-strategy.ec.europa.eu/en/news/final-report-high-level-expert-group-impact-digital-transformation-eu-labour-markets>

Smids, J., Nyholm, S., & Berkers, H. (2020). Robots in the workplace: a threat to—or opportunity for—meaningful work?. *Philosophy & Technology*, *33*(3), 503-522.

Tubaro (2020). The trainer, the verifier, the imitator: three ways in which platform workers support AI, *Big Data and Society,* 1-12

Zwolinski & Wertheimer (2016). Exploitation. *Stanford Encyclopedia of Philosophy*, <https://stanford.library.sydney.edu.au/archives/win2016/entries/exploitation/>

**WK12: Class presentation of group projects**

## Assessment

Assessment will include a weekly continuous assessments, including mini-MCQs (worth 10% overall) and online comments on class readings (worth 20% overall). The main assessment is a project (worth 70%) in which students create (in small groups or individually) a resource on one of the course topics. This resource is meant to communicate an ethical issue to a defined target audience, e.g. the general public or IT professionals. It can be a podcast, short video, infographic or blog post. This project will be presented together with a literature review and a reflection on the decisions you took to communicate the content effectively.

Guidance and/or rubrics for the assignments are available on Blackboard.

## Workload:

Lecture: 2 hours/weekly: 24 hours

Weekly readings and online comments: 2 hours/weekly: 24 hours

Group work and resource creation: 24 hours (Week 4-12)

Independent study: 24 hours

## Module resources

**Some recommended books and overview articles**

Markus Dubber, Frank Pasquale & Sunit Das (Eds.). (2021). *The Oxford Handbook of Ethics of AI*. Oxford Handbooks.

Virginia Eubanks 2018, *Automating Inequality*

Luciano Floridi (Ed.) 2010, *The Cambridge Handbook of Information and Computer Ethics*

Vincent Müller 2020, Ethics of Artificial Intelligence and Robotics, *Stanford Encyclopedia of Philosophy*, <https://plato.stanford.edu/entries/ethics-ai/>

Helen Nissenbaum 2009, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*

Shoshanna Zuboff 2019, *The Age of Surveillance Capitalism*

**Recommended Journals and conferences**

Big Data and Society

Science and Engineering Ethics

Ethics and Information Technology

AI and Ethics

AI and Society

FAccT Conference