Stanley Greenstein, Panagiotis Papapetrou
& Rami Mochaourab

Embedding Human Values into Artificial Intelligence (AI)[1]

1 Introduction

In the digital environment, the technologies that we design are value-laden whether we like it or not, whether intentional or unintentional and no matter how neutral we attempt to make these technologies. This is no different in the case of Artificial Intelligence (AI). The overall goal of this paper is to highlight the extent to which technology embodies human values based on ethics and morality, the extent to which the law prescribes that these human values be embodied in the technology that is created as well as the difficulty complying with these legal requirements

[1] This article originates from work performed and insights gained as part of the ongoing project EXTREMUM (Explainable and Ethical Machine Learning for Knowledge Discovery from Medical Data Sources). EXTREMUM is a Digital Futures sponsored project. More information about Digital Futures and the EXTREMUM project can be found at https://www.digitalfutures.kth.se/about/background/ and https://www.digitalfutures.kth.se/research/collaborative-projects/extremum/. The insights revolve around two points: 1) it has become commonplace to encourage representatives from different academic disciplines to work together to solve problems associated with modern digital technologies, however, this is easier said than done and as a legal scholar trained to work with legal tools, the ability to comprehend the mathematical and statistical language of the data scientists can be described as challenging to say the least; and 2) while the main question continually being asked by the data scientists is, 'will it work?', the main question continually being asked by legal practitioners is, 'is it good?', (adapted from a quote by Joseph Weizenbaum, *On the Impact of the Computer on Society*, Science, 176(4035), 609–614, 1972, pp. 611–612, in Friedman, Batya and Hendry, David G., *Value Sensitive Design: Shaping Technology With Moral Imagination*, Massachusetts Institute of Technology, 2019).

to the extent that from the technical perspective, many of these human values are mutually exclusive, meaning that promoting one human value is often at the expense of another competing human value. In other words, legal frameworks may list a whole range of human values that should be embedded in the technology developed, however, promoting one of these human values ultimately comes at the expense of another. The result is that those developing the technology are required to perform a balancing act to the extent that not all the mandated human values can co-exist in equal terms.

There has no doubt been a considerable hype surrounding AI during the recent past. Yet, there is still uncertainty concerning what the phenomenon entails. There are some that embrace the term AI while there are others that prefer the term machine learning. A common conception is also that AI is merely data and algorithms. This paper does not seek to describe what AI is nor does is seek to define it. Rather, its conceptual point of departure is to describe this technology, like many others have done before, by means of the metaphor of the "black box". It is a black box comprising various technologies, where data is fed into this black box and the black box proceeds to output data, usually in the form of knowledge on which decisions can then be based.

The idea of technology reflecting human values, such as ethical and moral values, is not new to scholars working in various academic disciplines that examine this idea. And even from the legal perspective, this idea is recognized by the legal regulator, one of the most evident examples of this being article 25 of the General Data Protection Regulation (GDPR) mandating data protection by design, in turn being based on the notion of Privacy-by-Design.[2]

There are many academic disciplines that deal with the notion of embedding values into technology. One of the academic areas of study in this regard is referred to as Value Sensitive Design (hereinafter referred to as VSD). A foundation upon which VSD resides is that of, '[c]reating computer technologies that – from an ethical position – we can and want

[2] Article 25, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016. On Privacy-by-Design, see Information and Privacy Commissioner of Ontario, *Introduction to PbD*, available at https://www.ipc.on.ca/english/privacy/introduction-to-pbd/.

to live with'.[3] The importance of considering what human values should be embedded in technology and at what point in time becomes critical to the extent that once the technology has been built, it is no longer possible to negotiate human values with machines. As Friedman states, VSD is important because, '[…] unlike with people with whom we can disagree about values, we cannot easily negotiate with the technology'.[4]

Indeed, the contemplation over the manner in which technology and society interacts is not restricted to a specific academic domain. The point of departure is that all human interaction occurs within an environment. The manner in which this environment is constructed is important, both from a symbolic point of view but also because changes to a physical environment influence behaviour. This is particularly well put by Winston Churchill, where, after the House of Commons was damaged by a bomb in 1941, it had to be decided whether to re-build it according to its previous design or whether to adopt a more modern design:

> Here is a very potent factor in our political life. The semicircular assembly, which appeals to political theorists, enables every individual or every group to move round the centre, adopting various shades of pink according as the weather changes … The party system is much favoured by the oblong form of the chamber. It is easy for the individual to move through those insensible gradations from Left to Right, but the act of crossing the Floor is one which requires serious attention.[5]

In this scenario, the political environment was reflected in the design of the chamber of Parliament and the symbolic value of forcing a person, wanting to change political party, to step to the other side of the chamber elevated the seriousness of such a political move. Churchill is also accredited with the saying,' [w]e shape our buildings; thereafter they shape

[3] Friedman, Batya, *Value Sensitive Design*, Interactions Volume 3 Issue 6 Nov./Dec. 1996, pp. 16–23 https://doi.org/10.1145/242485.242493, p. 17.

[4] Ibid., p. 21. Depending on future developments within AI and machine learning, this type of interaction with computer systems may be possible in the future, however, this issue remains outside the boundaries of this paper.

[5] Churchill, Winston S., *The Second World War, Volume V, Closing the Ring*, Cassell & Co, 1952, at p. 150 in Klang, Mathias, *Disruptive Technology – Effects of Technology Regulation on Democracy*, Gothenburg Studies in Informatics, Report 36, October, 2006, p. 1.

us', which too illuminates the notion that there is a strong bi-directional relationship between society and the technology that society creates.[6]

It is also important to recognize that human or moral values do not exist in a vacuum. For example, human values may be influenced by economic or political considerations. In the discipline of participatory design, a field that has close ties to VSD, reference is made to the Scandinavian context, where technologists and designers were working in a context with strong labour unions and co-determination laws, which in turn gave rise to a new approach to system design which sought to empower workers' sense of knowledge and a sense of work practice into the system design and development process.[7] This is an example of how a political context influences the human values that eventually are embedded in the design of technology. Another example is the effect that technological development has on the environment, where it is estimated that the electricity used to mine Bitcoin exceeds the consumption requirements of some countries.[8]

One can then reflect on the role of a legal practitioner working as part of the design team developing a new technology. It is argued that the legal practitioner working as part of a design team made up of data scientists should be, firstly to create an awareness of the interaction between human values and technological development and secondly, to promote the legal values that take on an increased relevance depending on the context within which the technological development is taking place.

The notion of contemplating which human values should be embodied into technology is not new. For example, Norbert Wiener, the American mathematician accredited with the establishment of the area of study called cybernetics, was interested in the field of computer technology, values and design. He authored the book *Cybernetics: Or Control and*

[6] World Scientific, available at https://www.worldscientific.com/doi/10.1142/9789813232501_0007, referenced in Friedman, Batya and Hendry, David G., *Value Sensitive Design: Shaping Technology With Moral Imagination*, Massachusetts Institute of Technology, 2019, p. 3.

[7] Friedman, Batya and Hendry, David G., *Value Sensitive Design: Shaping Technology with Moral Imagination*, Massachusetts Institute of Technology, 2019, p. 13.

[8] Aratani, Lauren, *Electricity needed to mine bitcoin is more than is used by 'entire countries'*, The Guardian, 2021, available at https://www.theguardian.com/technology/2021/feb/27/bitcoin-mining-electricity-use-environmental-impact.

Communication in the Animal and the Machine.[9] Shortly after the Second World War, Wiener started making references to the 'automatic age' or 'the second industrial revolution' as he put it and made references to the social and ethical challenges associated with these developments, and especially how information communication technology was bound to affect fundamental human rights.[10]

The underlying rationale to VSD rests on the relatively simple contention that human values shape technological development, these human values consequently become embedded in the technology itself and therefore the technology reflects the human values of the design team, whether intentionally or unintentionally.[11] There are a number of challenges for VSD, as acknowledged by Friedman and Hendry. A core element of VSD is that of morality and ethics. However within these disciplines that examine morality and ethics, e.g., philosophy, or legal theory, there are still ongoing academic discussions concerning various core concepts, discussions of the notion of justice within moral philosophy being one such example provided.[12] In other words, the above authors suggest that VSD is a philosophy that aims at continuing to deliberate the interaction between human values and technology, while moral philosophers, legal scholars and social scientists work these issues out.[13] VSD, therefore, is a philosophy that gives momentum to the idea of embedding human values into technology while other disciplines are bickering about theoretical issues. The above statement by Friedman and Hendry is slightly provocative to the extent that circumstances have changed and areas of law such as legal informatics are concerned with such topics.[14] Additionally,

[9] Wiener, Norbert, *Cybernetics: Or Control and Communication in the Animal and the Machine*, (Hermann & Cie) & Camb. Mass., MIT Press, 1948.

[10] Bynum, Terrell, *Norbert Wiener's Vision: The Impact of "the Automatic Age" on Our Moral Lives*, 2002, available at https://www.researchgate.net/publication/2537468_Norbert_Wiener%27s_Vision_The_Impact_of_the_Automatic_Age_on_Our_Moral_Lives.

[11] Friedman Batya and Kahn, Peter H., *Human Values, Ethics and Design*, University of Washington, pp. 1177–1201, available at https://depts.washington.edu/hints/publications/Human_Values_Ethics_Design.pdf, see also Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 32.

[12] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 7.

[13] Ibid.

[14] For a discussion of the discipline of legal informatics, see Greenstein Stanley, *Elevating Legal Informatics in the Digital Age*, in Petersson, Sonya (ed.), *Digital Human Sciences: New Objects- New Approaches*, Stockholm University, Stockholm University Press, 2020.

legal scholars are drawing attention to this theme. Bygrave, for example, referring to the widespread misconception that the logic of technology is beyond human control states that, '[…] this logic – as in the case of other technologies – embodies the values of its legal creators, and these values lay constraints on the technologies' use'.[15] On the other hand, this statement from Friedman and Hendry is inspirational and VSD remains an interesting design philosophy from within which to promote legal values (representing human values) into the design of technology.

It must be stressed that it is not the intention of this article to apply the VSD methodology to its fullest extent and exactly as promoted by the main proponents of this philosophy. Rather, VSD is used merely as a catalyst for inspiration based on the overall ideas that it portrays and also for its core stance concerning technology and values, i.e. technology embodies inbuilt human values which are inserted into the technology either intentionally or unintentionally. This article entails an adaption of VSD to the legal sphere, which in turn gives rise to some challenges, one being the fact that legal scholars for the most part are not that accustomed to making use of empirical studies, these being a central aspect of VSD as discussed below. In addition, for the most part, applying VSD from the legal perspective eradicates the need to go to such great lengths in order to identify which values are at play. This is already pre-determined to the extent that the values at play are those defined in the law. While it would be dangerous to assume that laws promote all ethical and moral values relevant to a specific context, the human values that are promoted by law serve as a good starting point as well as suffice to make a point.

The main argument that this paper seeks to make is the following: technology in the form of machine learning is comprised of values – mathematical and statistical values. Into this technology we seek to embody human values, based on considerations of morality and ethics and usually expressed in the natural language form. We are therefore dealing with two separate systems each comprising a different set of values displaying very different characteristics. If human values are to be embedded into technology and more specifically, if the human values mandated by the law are to be embedded into technology, then the human values re-

[15] Bygrave, Lee A., *Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions*, University of Oslo Faculty of Law Legal Studies, Research Paper Series, No. 2020-35, p. 5.

quire transforming into technical values, which is not an easy task and the illumination of which is at the core of this article.

2 Elevating Value Sensitive Design

First, this section examines the notion of values and more specifically provides some definitions of the concept of 'value'. This is relevant to the extent that it is human values and their embodiment into technology that is the focus of this paper. The notion of 'values' can be cause for confusion to the extent that values have a mathematical and data science connotation (technical values) as well as used in the sense of morality and ethics (human values). While this paper has its point of departure in the latter notion of values, the core point being argued is that the transformation of human values to mathematical values is not as straightforward as may seem, in turn requiring a balancing act. This section proceeds to investigate the academic discipline of VSD, essentially describing what it is and how it is can be applied.

2.1 Values

In examining a concept for the first time it can be useful to get an initial linguistic meaning or definition. Consequently, 'value' has been described as, 'something (such as a principle or quality) intrinsically valuable or desirable'.[16] It is also described as, '[p]rinciples or standards of behaviour; one's judgement of what is important in life.[17] Values have also been described as what is important to people in their lives, with a focus on ethics and morality.[18]

From the VSD perspective, the notion of what a value is has been described as follows:

> In some sense, we can say that any human activity reflects human values. I drink tea instead of soda. I recently attended a Cezanne exhibit instead of a ball game. I have personal values. We all do. But these are not the type of

[16] Merriam-Webster, *Value*, available at https://www.merriam-webster.com/dictionary/value.

[17] Lexico (Oxford), *Value*, https://www.lexico.com/definition/value.

[18] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 4.

> human values which this volume takes up. Rather, this volume is principally concerned with values that deal with human welfare and justice.[19]

The notion of what a value is, is also dealt with in the realm of legal theory:

> Values are whatever human beings hold to as the underpinning reasons behind more immediate reasons for acting, for approving action, and for preferring certain ways of acting and states of affairs to others. They are as such not necessarily themselves backed by further or ulterior reasons. This we express rather than explain by saying that, for us, something or other is 'good in itself'; whatever is good in itself is, for that person, an ultimate as distinct from a merely instrumental or derivative value. Hence arguments concerning what is of ultimate value cannot proceed by way of demonstration or proof.[20]

Besides examining values in themselves, the relationship between values and technology also encompasses a directional aspect in the manner that entails not merely the embedding of values in technology but also the fact that values and technology exert an influence upon each other. Nissenbaum, in examining the notion of how technology embodies values, highlights the unidirectional manner in which technology exerts an influence over society and which until now has been the predominant focus of scholars. She highlights two trends that exemplify this unidirectional approach: first, there is the situation of computers replacing humans in positions of responsibility, thereby affecting the extent to which society is able to hold humans accountable or responsible (without paying attention to the notion of responsibility as a value in itself); the second is where technological development forces us to re-examine the values themselves, e.g., how should we conceptualize privacy as a value in the light of a new technology. For Nissenbaum, the focus of study should be the opposite, namely, the direction 'from values to technology' and the manner in which values affect technology.[21] She states:

[19] Friedman, Batya, *Introduction*, in Friedman, Batya (ed.) *Human Values and the Design of Computer Technology*, Centre for the Study of Language and Information, 1997, p. 3.
[20] MacCormick, Neil, *H.L.A. Hart*, Stanford: Stanford Univ. Press, 1981, p. 48.
[21] Nissenbaum, Helen, *How Computer Systems Embody Values*, Computer, 2001, available at https://nissenbaum.tech.cornell.edu/papers/embodyvalues.pdf, p. 120.

> Humanists and social scientists can no longer bracket technical details—leaving them to someone else—as they focus on the social effects of technology. Fastidious attention to the before-and-after picture, however richly painted, is not enough. Sometimes a fine-grained understanding of systems—even down to gritty details of architecture, algorithm, code, and possibly the underlying physical characteristics—plays an essential part in describing and explaining the social, ethical, and political dimensions of new information technologies.[22]

It is often the case that when a public uproar erupts over a technological development, it is often the case that the technology has provoked a certain human value held dear by society. This line of argumentation is put forward by Nissenbaum who states that, 'the failure to meet technical criteria [does] not cause the public debate […] it was the controversial ways that these technologies engaged social, ethical and political values that did this.[23] Brownsword and Goodwin refer to certain concepts as 'boundary marking concepts'. These are concepts, they explain, that can be used in discussions about the desirability of certain technologies, and which mark the acceptable border of a technology in relation to morality.[24] In other words, these are concepts that draw the line for what is morally acceptable. Boundary marking concepts have certain distinguishing characteristics. First, they have the goal of being an instrument in determining the boundary of what technology is to be permitted. Second, they are not only concerned with prohibition. Third, a boundary marking concept may have within it a pre-defined notion of what is morally acceptable. For example, the notion that human dignity arises out of a religious belief in itself sets a boundary that is not open for negotiation. Fourth, boundary marking concepts do not exist in a vacuum, but rather reflect the norms of society, which themselves may not remain constant. A certain normative belief system will result in a certain boundary marking concept having a greater importance than another, e.g., those encompassing normative outlooks associated with a religious belief.[25]

[22] Nissenbaum, (n. 21), p. 121.
[23] Ibid., p. 118.
[24] Brownsword, Roger and Goodwin, Morag, *Law and the Technologies of the Twenty-First Century*, Cambridge University Press, 2012, p. 188.
[25] Ibid., p. 190.

2.2 Value Sensitive Design

Turning now to VSD itself, it can be described as a design philosophy that is one of the forebearers of the 'by-design' approach to technology. It is also described as a methodology or approach to the design of information and computer systems that came to the fore in the 1990's.[26] In its most basic form it is described in the following manner:

> Value Sensitive Design seeks to guide the shape of being with technology. It positions researchers, designers, engineers, policy makers, and anyone working at the intersection of technology and society to make insightful investigations into technological innovation in ways that foreground the well-being of human beings and the natural world. Specifically, it provides theory, method, and practice to account for human values in a principled and systematic manner throughout the technical design process.[27]

It is described as, 'a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process'.[28] The main thrust of VSD is to ensure that human values are embedded into the technological design process already from the design stage. The human values that one seeks to embed into the technology are those that have a moral characteristic, e.g. privacy, trust, accountability, honesty, freedom from bias and democracy, to mention but a few.[29] The spectrum of values addressed is relatively wide and may even include values associated with usability, conventions and personal taste.[30]

VSD embodies a number of commitments:

> the relationship between technology and human values is fundamentally interactional; analyses of both direct and indirect stakeholders; distinctions among designer values, values explicitly supported by the project and stake-

[26] Friedman, Batya, *Value Sensitive Design, Berkshire Encyclopedia of Human-Computer Interaction*, 2004, available at https://old.vsdesign.org/publications/pdf/friedman04vsd_encyclopedia.pdf, p. 769.

[27] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 3.

[28] Friedman, Batya, Kahn, Peter and Borning, Alan, Value Sensitive Design: Theory and Methods, UW CSE Technical Report, 2003, available at https://www.researchgate.net/publication/2551270_Value_Sensitive_Design_Theory_and_Methods, p. 1.

[29] Friedman, *Value Sensitive Design*, (n. 26), p. 769.

[30] Ibid., p. 769.

> holder values, individual, group, and societal levels of analysis; integrative and iterative conceptual, technical, and empirical investigations; co-evolution of technology and social structure; and a commitment to progress (not perfection).[31]

Also, important as far as VSD is concerned are the following: it is proactive in nature, it critically assesses human values as it carries them into the design process, it enlarges the scope of human values and it broadens and deepens the methodological approaches, drawing on anthropology, design, human-computer interaction, organizational studies, psychology, philosophy, sociology and software engineering, to mention but a few.[32]

At the core of the VSD design philosophy is a methodology that includes three distinct investigations, namely the investigations of a *conceptual*, *empirical* and *technical* nature, in an integrative and iterative manner. First, regarding the conceptual investigation, it can be said to, 'comprise philosophically informed analyses of the central constructs and issues under investigation […] how does the philosophical literature conceptualize certain values and provide criteria for their assessment and implementation? What values have standing? How should we engage in trade-offs among competing values on the design, implementation and use of information systems […]?[33] The conceptual investigation can be further described by focusing on the types of questions it seeks to address:

> Who are the stakeholders? What is likely to be at stake for people and other nonhuman stakeholders? What theoretical commitments and choice of conceptual framework, if any, are made? If the design team makes a commitment to a particular ethical or cultural framework to support principles reasoning, how would it be articulated and integrated into the design process? What values are likely to be implicated? How will values be framed and characterized? What conceptual models, if any, for operationalizing a given value or values will be employed? How will results from an empirical or technical investigation be integrated into the conceptual framework of

[31] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 4.

[32] Ibid.

[33] Friedman, *Value Sensitive Design*, (n. 26), p. 770. See also Friedman, Batya, Kahn, Peter and Borning, Alan, *Value Sensitive Design: Theory and Methods*, UW CSE Technical Report, 2003, available at https://www.researchgate.net/publication/2551270_Value_Sensitive_Design_Theory_and_Methods, p. 1.

> the project? What value-orientated criteria will be used to judge success of the design?[34]

Additionally, a characteristic of the conceptual investigation is its flexibility, ranging from 'armchair analyses' to more analytical types of investigations. As seen from the above statement, the identification of stakeholders is in focus. Here the conceptual analysis requires the identification of stakeholders, both direct but also indirect, the former category of stakeholder being those that interact directly with a system and the latter being those that are affected by the system, even though they do not use the system.[35] The identification of stakeholders can be important also from the legal point of view. For example, in 2011 the state of Nevada in the USA promulgated a law regulating autonomous vehicles, where rulemaking authority was granted to the Nevada Department of Transportation, which in turn consulted car manufacturers, Google, insurance companies and consumer groups.[36] This example highlights the fact that the stakeholders to a regulatory regime may change and in turn affect the traditional regulatory consultation processes.

One value that is important is that of autonomy. It can be described as, '[…] individuals who are self-determining, who are able to decide, plan, and act in ways that they believe will help them to achieve their goals and promote their values. People value autonomy because it is fundamental to human flourishing and self-development'.[37] However, there are limits to how much autonomy one can provide before autonomy actually starts diminishing. An example is where a product or system is developed for a task, say making presentations. A user will want access to the higher levels of the programme, e.g. how to make slides etc, but not the programme code. The more that the user needs to address at the programme code level, the more autonomy diminishes in that the user will not be able to achieve his or her goals. In this case, autonomy can be described as,

[34] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 32.

[35] Friedman, *Value Sensitive Design*, (n. 26), p. 770. An example provided in the literature is a system used in the health care context, where doctors and nurses would be the direct stakeholders and patients would be indirect stakeholders.

[36] Richards, Neil and Smart, William D., *How Should the Law Think About Robots?*, in Calo, Ryan, Froomkin, Michael A., and Kerr, Ian (eds.), *Robot Law*, Edward Elgar, 2016, p. 12.

[37] Friedman, *Value Sensitive Design*, (n. 3), pp. 17–18.

'when users are given control over the right things at the right time'.[38] Autonomy can therefore be seen in terms of system capability where autonomy can be undermined when the computer system does not provide the user with the necessary technological capability to realize his or her goals.[39] Another value identified in technology is that of bias, a definition being that:

> We say that a computer technology is biased if it systematically and unfairly discriminates against certain individuals or groups of individuals in favour of others. A technology discriminates unfairly if it denies an opportunity or a good, or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate.[40]

The second investigation as part of the VSD methodology is the empirical investigation, which essentially validates and potentially expands the values identified in the conceptual investigation. In other words, while the conceptual investigation may assume a number of relevant values in a specified context, the empirical investigation can refine these values and also confirm the assumptions made in the conceptual investigation. The empirical investigation is required to the extent that the conceptual investigation, 'can only go so far' and that the human context in which the technology operates too needs investigation.[41] Put another way, the reason for applying the empirical investigation is purported to be the limitations connected to applying only a conceptual investigation.[42] Once again, the questions to be asked as part of the empirical investigation are the following:

> How do stakeholders apprehend individual values in the sociotechnical context? How do stakeholders prioritize competing values or otherwise envision resolution of value tensions? Are there differences between espoused practice (what people say) compared with actual practice (what people do)? […] [w]hat are organizations' motivations, methods of training and dissemination, reward structures, and economic incentives?[43]

[38] Friedman, *Value Sensitive Design*, (n. 3), p. 18.
[39] Ibid., p. 18.
[40] Ibid.
[41] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 33.
[42] Ibid.
[43] Ibid., pp. 33–34.

Finally, the technical investigation focuses on the design and performance of the technology itself, where the assumption is that, '[…] technologies in general, and information and computer technologies in particular, provide value "suitabilities" that follow from the properties of the technology […] a given technology is more suitable for certain activities and more readily supports certain values while rendering other activities and values more difficult to realize'.[44] Two additional aspects characterise the technical investigation, namely how technical properties support or hinder human values and secondly the proactive aspect whereby the design properties can be promoted with the intention of promoting the human values identified in the conceptual or empirical investigation.[45] In addressing the technical investigation, the following questions are addressed:

> What features of a technical infrastructure enable, hinder, or even foreclose certain kinds of designs for supporting human activity? How do policies, laws, or regulations create opportunities or constrain options for technological development?[46]

Consequently, VSD design embodies these three investigations that should occur in an integrative and iterative manner. In other words, all these separate investigations in effect influence each other. A description of the VSD methodology reveals that a minimal focus is placed on legal consideration. For example, the questions addressing the technical investigation do consider the role of laws and regulations. However, it is argued that as the awareness of how technology embodies values increases within the legal profession and more importantly within the realm of the legal regulator, so too are the legal requirements increasing that mandate the embedding of legal values (representing human values) in technology. The next section illuminates the EXTREMUM project and the extent to which legal values mandated by law are embedded into the design process of the technological development.

[44] Friedman, *Value Sensitive Design*, (n. 26), p. 770.
[45] Ibid., p. 770.
[46] Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, (n. 7), p. 34.

3 Project EXTREMUM

The EXTREMUM project can in simple terms be described as a machine learning initiative whereby useful knowledge is extracted from databases comprising medical data. The knowledge that is sought relates to the adverse effect of certain prescription drugs in order that the adverse effects can be predicted and prevented. The same applies to the detection and predictive treatment of patients in relation to cardiovascular diseases. The ultimate goal of the project is to develop a prototype system that can be used to achieve the above insights from health data and is best explained by an extract from the project web site:

> to develop a novel platform for learning from complex medical data sources with focus on two healthcare application areas: adverse drug event detection and early detection and treatment of cardiovascular diseases [it] will present a new framework for data management and analysis of the integration of data, methods for machine learning as well as ethical issues related to predictive models. The fundamental breakthrough of this project is to establish a novel knowledge management and discovery framework for medical data sources. The outcome will be a set of methods and tools for integrating complex medical data sources, a set of predictive models for learning from these sources with emphasis on interpretability and explanatory features, and simultaneously focusing on maintaining ethical integrity in the underlying decision mechanisms that rule the machine learning.[47]

From the above text it becomes apparent that the tools used to extract knowledge from the medical data are complex machine learning algorithms and models. What the VSD philosophy and methodology shows us is that these algorithms and models are not value-neutral to the extent that they include social, ethical and moral values – human values. The next consideration entails which values should be integrated or embedded into these machine learning algorithms and models. The point of departure is naturally that there is a wide spectrum of values that could potentially be relevant, depending on which stakeholders you address as well as what context one is considering. Which values should therefore be included for consideration? In order to simplify matters, and from the legal perspective, the answer to this question is rather obvious – it is the human values as mandated and promoted in the formally promulgated

[47] Project EXTREMUM, https://www.digitalfutures.kth.se/research/collaborative-projects/extremum/.

laws that regulate a relevant context. Once again, not all laws can be consulted and not all human values addressed in these laws can be taken into account. For example, it may not only be formal legal instruments that promote values but other regulatory instruments that reside under the banner 'soft law' may also be relevant.[48] However, employing the VSD technical investigation, and also based on legal experience, one soon recognizes which legal frameworks are more relevant in the given context. For example, a project working with personal data in the form of health data naturally calls into consideration the General Data Protection Regulation (GDPR) and consequently the values this legal instrument promotes.[49] Even within the GDPR a wide range of human values may exist. The legal values addressed in the next section have been selected in order to illuminate the main goal of this paper and have therefore been chosen not only because of their relevance but also because of the pedagogical value that a discussion of these values promotes.

3.1 Identified Legal Values

Three human values have been identified as being relevant to the extent that they are mandated by the GDPR. These three values are *explainability*, *privacy* and *accuracy*. An in-depth analysis of these three concepts or values remains beyond the boundaries of this paper and their choice is merely illustrative of the fact that human values are mandated by legal instruments when designing technology. There are many human values promoted by the GDPR, e.g., autonomy, personal integrity and dignity, to name but a few. The three values of explainability, privacy and accuracy have been chosen as they have one thing in common – that is, echoing Friedman above, they promote human welfare and justice. This said, a very brief explanation of the above human values follows.

A central principle of the GDPR is that the data subject be granted information concerning the processing of personal data. More specifically, Article 13(2)(f), 14(2)(g) and 15(1)(h) regulate the provision of in-

[48] Here an example of a 'soft law' code is the *Ethics Guidelines for Trustworthy AI* by the European Commission High-Level Expert Working Group on Artificial Intelligence, available at https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

[49] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016.

formation in relation to Articles 22 concerning automated decisions. In addition, Recital 63 is relevant to the extent that the data subject should have a right of access to, 'the logic involved in any automatic personal data processing and, at least when based on profiling'. In relation to Article 22, a reference to explainability is found in Recital 71, which states that the data subject has, 'the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision'. Whether the explanation is formed as a right is beyond the bounds of this paper.[50] What is noteworthy for the purposes of this paper is that explainability exists as a value that is promoted by the GDPR. And the justification for enshrining explainability is illuminated by Bygrave in his reference to opaque machine learning decisional systems and the human interest in 'cognitive sovereignty', stating that, '[t]he interest is foundational to the normative justification for requiring explicability of machine processes'.[51] It is argued that there is still considerable discussion concerning what explainability actually entails and many questions arise: what is meant by this explainability? Explainability for which stakeholder? Is it explainability for data scientists or data subjects? And is explainability even attainable in the era of deep neural networks, the inner workings of which go beyond the cognitive ability of human beings?

The questions are complex and the answers elusive. However, the intention of this paper's emersion in the notion of values such as explainabilty is echoed by Brkan and Bonnet:

> The GDPR is thus becoming increasingly important also for XAI researchers and algorithm developers, since the introduction of the legal requirement for understanding the logic and hence explanation of algorithmic de-

[50] For an in-depth discussion surrounding explainability as a right, see Goodman B and Flaxman S, *European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"*, ICML Workshop on Human Interpretability in Machine Learning, arXiv:1606.08813**,** AI Magazine, Vol 38, No 3, 2017, Wachter S, Mittelstadt B, and Floridi L, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, in International Data Privacy Law, Volume 7, Issue 2, May 2017, pp. 76–99 and Selbst, A and Powles J, Meaningful Information and the Right to Explanation, in International Data Privacy Law, Vol. 7, No. 4, 2017, pp. 233–242.

[51] Bygrave, *Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions*, (n. 15), p. 8.

> cisions entails also the requirement to guarantee the practical feasibility of such explanations from a computer science perspective.[52]

The value of privacy is promoted by the GDPR in its entirety. However, one of the main Articles in which it finds expression is Article 25, entitled 'Data protection by design and by default'. The connection between Article 25 and the notion of privacy is indisputable. In the words of Bygrave, '[a]rticle 25 springs out of a policy discourse that commonly goes under the nomenclature 'Privacy by Design' ('PbD').[53] Also, in 2010 the 32nd International Conference of Data Protection and Privacy Commissioners passed a resolution to the effect that Privacy by Design was a fundamental part of fundamental privacy protection.[54] Ann Cavoukian is credited with coining the notion 'Privacy by Design', which is briefly described as, '… an approach to protecting privacy by embedding it into the design specifications of technologies, business practices, and physical infrastructures. That means building in privacy up front – right into the design specifications and architecture of new systems and processes'.[55] The main rationale to data protection by design is that privacy-related interests receive serious consideration throughout the entire lifecycle of information systems development and not just at the end.[56]

Article 25(1) states:

> Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary

[52] Brkan, M., and Bonnet, G., *Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas*, European Journal of Risk Regulation, Vol. 11, Issue 1, March 2020, pp. 18–50, p. 19.

[53] Bygrave, Lee A., *Data protection by design and by default*, in Kuner Christopher, Bygrave Lee A. and Docksey Christopher (eds.), *The EU General Data Protection Regulation (GDPR)*, Oxford University Press, UK, 2020, p. 571.

[54] Ibid., p. 571.

[55] Information and Privacy Commissioner of Ontario, *Introduction to PbD*, available at https://www.ipc.on.ca/english/privacy/introduction-to-pbd/ (last accessed on 2016-03-24).

[56] Bygrave, *Data protection by design and by default*, (n. 53), p. 571.

> safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

For the purposes of this paper, it is sufficient to note that privacy is an inherent value protected and promoted by the GDPR and that it finds expression throughout the GDPR but also very concretely in Article 25 GDPR.

The GDPR incorporates a number of data protection principles in Article 5, and in Article 5(1)(d) accuracy is mentioned as a principle. The rationale for incorporating accuracy as a principle is brought to the fore and expanded upon by the Article 29 Data Protection Working ('Working Party') Party.[57] The Working party specifically mentions accuracy in relation to profiling and that it should be taken into account during the collection of data, the analysis of data, the building of a profile and the application of a profile.[58] The rationale for the need for accuracy is provided by the Working Group:

> If the data used in an automated decision-making or profiling process is inaccurate, any resultant decision or profile will be flawed. Decisions may be made on the basis of outdated data or the incorrect interpretation of external data. Inaccuracies may lead to inappropriate predictions or statements about, for example, someone's health, credit or insurance risk.[59]

Finally, it is argued that profiling may include an element of prediction, which in turn can result in inaccuracies if the underlying data is incorrect.[60] Here is can be argued that accuracy is necessary in order that an algorithm, creating a profile of an individual, paints as true a picture as possible of that individual. Accuracy can therefore be seen as a component necessary to gauge a person's reputation.[61] In this sense it is a human value worth preserving and promoting.

[57] Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, Adopted on 3 October 2017 available at wp251rev_01_en_A754F3E1-FB46-9E76-C0A919864E4B6641_49826.pdf.

[58] Ibid., p. 12.

[59] Ibid.

[60] Ibid., p. 17.

[61] For a discussion on reputation, see Greenstein, Stanley, *Our Humanity Exposed: Predictive Modelling in a Legal Context*, Dissertation, Stockholm University, 2017, available at http://www.diva-portal.org/smash/record.jsf?dswid=5270&pid=diva2%3A1088890.

The next section entails a closer examination of how the values of explainability, privacy and accuracy to some extent compete with each other but also complement each other.

3.2 The Trade-Off in Transforming Legal Values to Technical Values

This section draws on experiences from the EXTREMUM project in order to illustrate the challenges associated with transforming human values or legal values expressed in the natural language form into technical values represented by machines. The process of transforming human values into technical values is depicted in Figure 1 below. It is argued that the three legal values mentioned above, namely, explainability, privacy and accuracy, can be categorized into three separate techniques: the first is the learning of a machine learning classifier that can accurately diagnose patients based on historical patient data, the second is preserving the privacy of patients' data, and the third is providing explanations to diagnoses made by the developed machine learning models.[62]

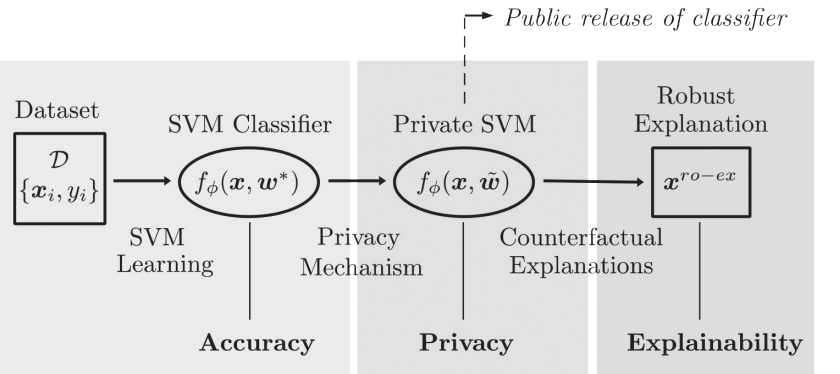

Figure 1 is a representation of the machine learning techniques performed in order to gain the insights from medical data in accordance with the goals of the EXTREMUM project.

In Figure 1, a dataset containing sensitive data with many patients' historical health records and respective diagnosis, represented by 'D', is se-

[62] Mochaourab, R, Sinha, S., Greenstein, S. and Papapetrou, P., *Robust Counterfactual Explanations for Privacy-Preserving SVM*, International Conference on Machine Learning (ICML 2021), Workshop on Socially Responsible Machine Learning, Jul. 2021.

lected. It is assumed that the dataset is securely stored within the confines of the hospital's technical infrastructure without public access to its entries. Within the secure confines of the hospital, the dataset is employed to train a machine learning classifier, in this case a Support Vector Machine (SVM),[63] that would predict the diagnosis of future patients based on the available historical data. The objective of SVM learning is to achieve the highest possible prediction accuracy and accordingly perform correct diagnosis for most future patients, i.e. as many patients as is technically possible. Hence, the value of accuracy is the prime goal.

The functionality of the SVM classifier depends on a set of parameters which are determined using the patients' health records in the dataset. Hence, any public accessibility to the trained SVM classifier parameters may lead to privacy breaches if an adversary manages to reconstruct the patients' dataset using the classifier parameters. Therefore, we need to ensure that the privacy of the persons in the dataset is preserved before publicly releasing the classifier. The privacy mechanism used here guarantees differential privacy, which is a privacy mechanism that incorporates a tuneable degree of uncertainty about the actual presence of any entry in the dataset (also referred to as 'noise').[64] This uncertainty is achieved through random perturbation of the SVM classifier parameters.[65] Consequently, the private version of the SVM classifier can be made publicly available with potential utilization in various contexts, e.g., in many different hospitals.

The benefit in guaranteeing privacy comes at the cost of reduced classifier accuracy. In other words, these two technical values are mutually exclusive to the extent that increasing one of them decreases the other. Figure 2(a) below illustrates the differences between the optimal SVM classifier and its private version. Firstly, two categories of patients are represented. The category of 'healthy patients' is represented by the circles and the category 'unhealthy patients' is represented by the plus signs. The two axes in the figure represent two features of the data, i.e., two attributes of the patients' health records. Examples of features can be smoking,

[63] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference and prediction*. Springer Series in Statistics. Springer-Verlag New York, 2nd edition, 2009.

[64] Dwork, C. and Roth, A. *The algorithmic foundations of differential privacy*. Found. Trends Theor. Comput. Sci., 9(3–4):211–407, August 2014. ISSN 1551-305X.

[65] Pertubation can be described as the adding of noise to data in order to enhance confidentiality.

amount of exercise, genetic indicators or previous health issues, to name a few. The optimal SVM decision boundary separates the two of patients classes with a straight line which maximizes the widths of the margins. Accordingly, the optimal SVM decision boundary is robust to any small changes in the data since it is furthest away from both sets of points (i.e. both circles and plus signs). On the other hand, the private SVM boundary (indicated with the broken line), which is a randomly perturbed version of the optimal SVM, is clearly less robust than the optimal SVM boundary (its distance to the data sets has diminished and is therefore not as robust). This means that its accuracy in classifying future patient cases may on average be lower than that of the optimal SVM while on the other hand it enhances privacy. In this way, the striving after privacy in technical terms comes at the expense of accuracy.

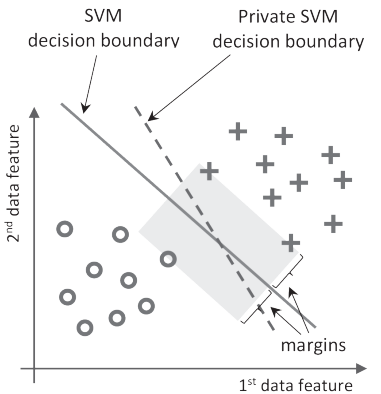

Figure 2(a)

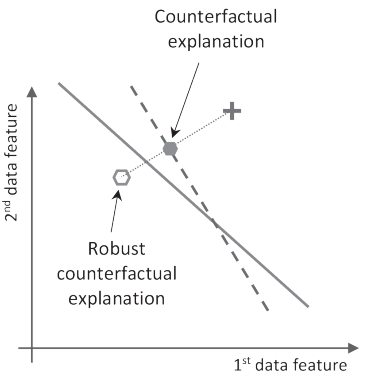

Figure 2(b)

Figure 2 (a) illustrates the machine learning classifiers and the differences between the optimal and private SVM. Figure 2 (b) shows a counterfactual explanation for a single data point as well as its robust version.

Figure 1, also proceeds to consider the explainability of classifications made by the private SVM (privacy preserving SVM). This can be explained by a hypothetical example depicting the patient and medical practitioner context. On informing a patient about the diagnosis provided by the classifier, it becomes possible to quantify the least necessary changes to the patient's data that would lead to another diagnosis. For example, this could entail informing a patient of what is required for him or her to move from the category of ill patients to the category of healthy

patients. This can be related to Figure 2(b), where the changes lead to a given patient's data point on one side of the SVM decision boundary moving to the opposite side of the SVM decision boundary. Subsequently, the patient perceives a contrastive example of related data which helps in explaining the diagnosis. Such types of explanations are called counterfactual explanations and are especially useful when the changes in the data are actionable, i.e., the patient is able to perform certain tasks to change the outcome of the diagnosis.[66]

Providing valid counterfactual explanations for the privacy preserving SVM classifier is a challenging task due to the introduced perturbations to the optimal classifier parameters. Observe that the optimal SVM classifier is unknown since only the private SVM is publicly released. The necessary changes to the patient's data that lead to a different diagnosis according to the private SVM classifier may still give the same original diagnosis according to the optimal SVM classifier, as is shown in Figure 2(b). Addressing this issue requires studying robust counterfactual explanations that consider the extent of perturbations required by the privacy mechanism.[67] Generating larger perturbations to achieve larger levels of differential privacy would essentially require larger changes in the patient's data for counterfactual explanations. This is illustrated in Figure 2(b) where the robust counterfactual explanation is further away on the other side of the boundary to make sure that it is correctly classified according to the optimal SVM decision boundary. In other words, these larger changes would guarantee a desired level of confidence that we predict a different diagnosis using the unknown optimal SVM classifier. Hence, as a summary, guaranteeing a desired level of differential privacy diminishes the classifier accuracy and consequently increases the required changes in counterfactual explanations to meet a certain level of confidence in validity of the explanations.

[66] Wachter, Sandra, Mittelstadt, Brent C. R., *Counterfactual explanations without opening the black box: Automated decisions and the GDPR*. Harvard Journal of Law & Technology, 31(2), 2018.

[67] In this regard reference is made to footnote 62 above.

4 Conclusions

This paper began with the argument that whether we like it or not, the technology we create has embedded within it human values, more specifically social, ethical and moral values. There may be varying interpretations as to what a value is and where the boundaries lie as far as values are concerned, but for the sake of simplicity, a value can be said to be a good in itself. The paper then proceeded to illuminate the philosophy of VSD, which places a large emphasis on identifying the human values associated with technological development. However, from the purely legal perspective, the human values inherent in regulatory instruments provide a natural point of departure for a discussion of values.

Using knowledge from research within the EXTREMUM project, the main argument put forward in this article is that having identified the human values relevant in relation to a particular technology and social context, it may not be a straightforward issue of transposing these into the language of data science. Challenges include the fact that the values expressed in laws have not undergone a balancing process. In other words, the GDPR refers to explainability, privacy and accuracy but it does not consider the difficulties in embedding these values into the technology. However, these difficulties become apparent when the process of the translation of the values from natural language to the language of data science begins. It soon becomes apparent that in technical terms these human values are mutually exclusive – promoting one will invariably occur at the expense of another – which in turn leads to the next problem of having to balance these competing human values against one another as they are transformed into their mathematical equivalents. This is depicted in the EXTREMUM project where the human values of privacy and accuracy are pitted against each other as they are transformed into their mathematical and statistical equivalents, or put another way, into the rules of data science. Adding to the picture is the value of explainability which adds a layer of complexity.

Much attention is given to the fact that cross-disciplinary work is required to address the challenges of modern technological development. However, this is easier said than done, as depicted by the findings highlighted in this paper. One issue that comes to the surface is extremely important. The data scientist's reflex is to focus on the value of accuracy, which is nothing strange as his or her main focus is to develop technology that works in a manner that is as accurate as possible. However,

professionals from other disciplines such as the law, bring other insights to the table regarding human values and legal demands, which in turn will focus attention on other values in addition to accuracy, privacy being a prime example. The fact that values such as privacy are mandated by the law essentially create a dilemma for the data scientist. Naturally, the technology should be as accurate as possible, but adhering to the law may mean that part of the accuracy must be sacrificed for the sake of privacy – not because we want to, but because we have to – if we want to follow the law, that is. The follow-up question is extremely interesting, namely, if we then are mandated by law to insert privacy into technology, how much privacy is enough privacy according to the law? This in turn raises additional questions, e.g. how to quantify privacy and what level of privacy is demanded by the law? These are questions that remain outside the boundaries of this paper but that will hopefully be addresses in future works and fora.

It is argued that the above experiment brings worthwhile insights from various perspectives: it can be worthwhile from the legislative technique's perspective, i.e. in relation to how we can better produce laws and other sources of law; it is a valuable insight for legal practitioners called upon to ensure that human values and more specifically legal values are embedded into the technology we create and finally it creates awareness surrounding the insight that the technology we create reflects the human values we embrace.