AI Ethics 'From Below'

Candidate 1061912

Trinity Term, 2022

Word Count: 14,948

Thesis submitted in partial fulfilment of the requirement for the degree of MSc in the Social Science of the Internet at the Oxford Internet Institute at the University of Oxford.

Acknowledgements

I would like to express my gratitude to my supervisor for her patience and insightfulness. It was her work which inspired me to pursue AI ethics, and her unwavering support that gave me the confidence to contribute to such a vital field.

I would also like to thank my friends David Wong, Mneera Saud, Maya Sherman and Jeffrey Gan for the inspiration, joy and intrigue they persistently invoke. Finally, I would like to thank my parents, Charles and Sreela, for their love and support for everything I choose to do.

Abstract

The contemporary field of AI ethics consistently views its subject matter from a 'top-down' perspective, providing guidance for those with agency over how AI functions (usually engineers, designers, legislators and regulators). However, to the best of this candidate's knowledge, the discipline lacks any systematic attempts to generate AI ethics for those who use AI every day and live with the consequences of its proliferation. As such, the field disregards laypeople's agency over how AI is experienced.

This study seeks to generate ethical frameworks, centred around the values of autonomy and transparency, to support individuals in their interactions with AI. In order to do this, it uses analysis of philosophical concepts from scholars such as Dworkin and Zagzebski, inversions of key concepts from the AI ethics literature, and interviews with nine AI ethicists, to triangulate the best conceptualisations and operationalisations for normal people to employ. By using the presented frameworks, individuals can improve their knowledge and understanding of AI, and protect and enhance their autonomy around AI, leading to real benefits in quality of life.

Table of Contents

1. Introduction

Artificial intelligence (AI) has permeated into the lives of everyday individuals. It mediates their interactions and reforms the environments in which they live (Calvo et al., 2020). In response to the scale and scope of AI's effects, modern AI ethics has developed to mitigate its risks and promote its benefits (Taddeo & Floridi, 2018). A dense literature has emerged, covering the prevention of bias and discrimination; the preservation of autonomy and privacy; the establishment of meaningful transparency; the promotion of responsibility; and the advancement of justice; among other issues (Mittelstadt et al., 2016).

Different AI ethicists conceptualize AI differently. However, for the purposes of ethical analysis one of the best definitions of AI is offered by Taddeo & Floridi as:

> 'a growing resource of interactive, autonomous, self-learning agency, which enables computational artifacts to perform tasks that otherwise would require human intelligence to be executed successfully' (Taddeo & Floridi, 2018)

This is a useful socio-technical perspective. It emphasizes AI's instrumentality, or its ability to perform set tasks, rather than 'think' or 'take moral responsibility' itself (Dignum, 2017). It also highlights its untapped nature, or the extent to which individuals have hardly begun to drain the "reservoir" of "smart agency" and exhaust its applications (Floridi et al., 2018).

AI ethicists have inspired and implemented changes which have reduced the extent of AI's ill effects and capitalised on its potential. However, they have only done so from a single direction: the 'top down'. Indeed, since almost all current AI ethics resources aim to change only the behaviour of those with agency over how AI functions, the field is characterized by a fundamental asymmetry.

To illustrate the importance of the individual's agency in counteracting the unethical effects of AI, let us consider a benign example. An individual's music taste is being manipulated towards certain genres by their Spotify 'Discover Weekly' playlist. This is ethically problematic because it erodes their sense of autonomy over this element of their lives and potentially harms their experience of a beloved hobby. In this example, the individual can do a number of things to mitigate the risk of experiencing the ill effects of the system's AI. They can *recognize* the problem. They can *change* their use

of that playlist. They can *learn* the specifics of how the algorithm manipulates their taste in an effort to counteract it actively. They can *create* a new account to reset their preferences. They can *adjust* their privacy settings. They can *defer* taste-making power to newspapers and magazines which are more transparent in their biases and more selective in their recommendations. In short, they can take advantage of the philosophical observation that the subject of an action has agency over how that action is experienced, and how they respond to its occurrence (Schlosser, 2019). The importance of this agency is even more obvious regarding the promotion of life-enhancing AI applications; here, the relevant systems have to be actively *understood*, *adopted* and *used well*, in order to have a positive impact.

Despite its potency, the idea that individuals should play a role in mitigating AI's unethical impacts and promoting its benefits in their everyday lives, is foreign to the current field of AI ethics. Although this is obvious throughout the mainstream literature, it is made particularly clear from the recommendations of the European Institute for Science, Media and Democracy group, 'AI4People'. Composed of thirteen leading AI ethicists, this group made 20 recommendations for "next steps" to advance towards a world of more ethical AI, in 2018 (Floridi et al., 2018). These "next steps" aimed "to assess, to develop, to incentivize, and to support" "good AI" (Floridi et al., 2018). All of the group's proposed actions could only be undertaken by experts with agency over AI's function; none of them could be undertaken by the layperson.

Admittedly, the idea that individual agency affects interactions with technology is not foreign to adjacent literatures. For instance, there is an active academic discourse surrounding practical steps towards the protection of personal data (Cherry & LaRock, 2014). Moreover, the 'Human-Computer Interactions' (HCI) literature provides useful analysis of practical user experiences (Karray et al., 2017). However, these literatures do not perfectly align with the concerns of AI ethics or the values it seeks to protect. This is reasonable, as AI presents certain heightened risks in specific areas of an individual's life.

This thesis aims to guide people who interact with AI in expressions of individual agency. It provides justifications for thinking in terms of certain ethical ideas, and practical operationalizations of them. If individuals implement this project's ethical frameworks, they will make better decisions, based on better ideas, when engaging with

AI systems. Although it seeks to correct a significant imbalance in the AI ethics literature, it will not jettison the field's findings. In this way, the thesis lives up to its name. Indeed, the term 'from below' is taken from the mid-20th century French historians who formed the 'Annales school'. The Annales argued that political, social and economic history should not simply be concerned with kings, courts and councils, but with everyday individuals and their lives (Burke, 2015). They did not abandon the focus of the field, they simply inverted it. This project will do the same.

This thesis is structured as follows. The next chapter, entitled 'literature review', explains the theory used by this thesis. Chapter three, 'methodology', explains its methods, which involved nine interviews with AI ethicists. Chapter four provides a justification and explanation of two ethical frameworks which individuals can use in their interactions with AI. Chapter five outlines the limitations of this thesis, and chapter six concludes.

2. Literature Review

2.1 Which Values, For Whom?

AI ethics frameworks for the individual should focus on values which are both important to everyday life and feasible to operationalize. Not all individuals are concerned with grand societal improvements. Moreover, without power to redesign AI, an understanding of highly technical issues is difficult to translate into practical action. However, all individuals can effectively conceptualize and operationalize privacy, autonomy, and transparency. They can avoid surveillance as far as possible and act prudently when volunteering sensitive information to preserve their privacy; reflect upon and change their behaviour to regain autonomy; and 'make' AI systems more transparent by learning about them. Due to considerations of length and given the extensive academic and popular discourse on privacy, this thesis will focus upon autonomy and transparency alone.

This thesis' findings cannot apply to everyone. AI's effects are heterogenous, often varying by social settings (Hagerty & Rubinov, 2019); moreover, ethical beliefs vary by time and place (Skinner, 1969). Theoretically, this framework is useful for those who genuinely *value* transparency and autonomy and are *able* to operationalize them. Those whose views are totally philosophically unaligned with Western philosophy and those whose technological activities are totally coercive in character will be effectively excluded. There may also be practical constraints affecting the audience for AI ethics 'from below', which will be considered in chapter 5 on limitations.

2.2 Autonomy

Autonomy is highly valued in almost every ethical school. Philosophers have argued that it is "one of the elements of well-being" (Mill, 1859) and "something whose presence in our lives makes them go better in itself" (Sumner, 1996). In 'Welfare, Happiness and Ethics', Sumner argues that the alignment of choices with values and aims, and the 'authentic' self-endorsement of one's way of life which follows, is vital to welfare (Sumner, 1996). This is congruent with liberal positions, which suggest that individual autonomy must be preserved, in particular via non-interventionism by exterior institutions, because action must be aligned with one's own desires, which will vary from person to person (Hayek, 1960). Other philosophers suggest that autonomy's facilitation of long-term "self-development" (Christman, 2020), concurrently imbues individuals with a sense of personal responsibility which is itself a gratifying and

motivating force (Amartya Sen, in Sumner, 1996). All of these ideas are reinforced by empirical psychology. Indeed, 'self-development theory' (SDT), has replicated experiments showing that autonomy, competence, and relatedness are vital to "self-motivation and psychological wellbeing" (Calvo, 2020), as well as "performance", "persistence", "creativity" and "effective problem solving" (Varshney, 2020). SDT even explicitly places autonomy above competence and relatedness, as the most important feeling humans can experience (Varshney, 2020).

The most renowned philosopher of autonomy, Immanuel Kant, extends autonomy's importance beyond welfare. Kant argues that autonomy is "a necessary presupposition of all morality" (Kant, 1785). This is because "autonomy of the will" is "a species of [rationally-motivated] causality in living beings" (Kant, 1785); without it, individuals succumb to their base inclinations and fail to consistently act in alignment with rational moral premises like Kant's categorical imperative (Hill, 2013). Although they conceptualize its importance differently, Kant is not alone in presenting autonomy as vital to morality; Rawls, Scanlon, Wolff, and many others base their moral theories upon it (Dworkin, 1988). Kant's inverse of autonomy is 'heteronomy'. This is a state in which an individual's will is given laws, or controlled, by unwarranted forces, like their fleeting desires or the actions of others (Kant, 1785). Given the importance of avoiding heteronomy for moral action and the use of reason, Kant argues that autonomy is the basis for "human dignity" (Kant, 1785).

Despite their agreement on its importance, philosophers disagree about how autonomy should be conceptualized. The word itself derives from Archaic Greece, where "'autos' (self) and 'nomos' (rule or law)" were combined to describe city states where citizens legislated for themselves, rather than being controlled by conquerors (Dworkin, 1988). This premise has influenced later conceptualizations.

Most notably, for Kant, an autonomous will is one which is "law unto itself", i.e., bound by its own *reason* (Kant, 1785). Although this is an excellent basic idea of autonomy, the Kantian conception is tied up in metaphysical arguments and specifically deontological theories. Kant argues that, if it is believed that we must act on moral duties, there must be a "fundamental synthetic a priori proposition" (i.e., a truth which is fundamentally accurate, independently of its own premises and independently of experience), that a rational agent would follow the categorical imperative, i.e., that

individuals must "act only on that maxim by which you can at the same time will that it should become a universal law" (Kant, 1785; Hill, 2013). This definition of rationality, which is inherently bound to Kantian autonomy, is excessively constrictive. It is also excessively complex, and therefore confusing for those without a background in philosophy.

Later philosophers of autonomy have suggested other conceptualizations. However, many are unrealistic or incomprehensible. Rousseau's idea of 'moral liberty' is very similar to Kantian autonomy but can only be realized when citizens literally prescribe themselves laws via a *perfect* state which always accepts the 'general will' (Hill, 2013). Hare suggests that autonomy should reject narrow ideas of reason and argues that "principles commonly thought to be immoral" can be taken up by autonomous persons (Hill, 2013). This premise falls apart at the extremes; someone who believes that throwing themselves off a cliff is reasonable, would not normally be described as totally autonomous. Sartre's conception of freedom is similar to autonomy, but is framed as a curse; indeed, for Sartre, individuals are "condemned to be free", as they constantly remake themselves in each moment of their existence, confronting what they believe and what they should do in an unending low-level existential crisis (Sartre, 1956). This is certainly not the autonomy which should be protected.

Modern scholars, often in the legal literature, conceptualize autonomy more understandably as a *right*, a *capacity*, or a *state*. It is the right to make choices without interference; the capacity to make independent and reflectively-motivated decisions; or the state of being in control over how you live (Hill, 2013). All of these ideas are sensible, realistic, and compatible with one another. Yet they are excessively simple. They suffice for lawmakers, and indeed mainstream AI ethicists, who make quick judgements about heterogenized 'average' users or consumers. However, the individual requires a definition of more depth, simply because they are able to analyse their particular selves in more detail. They require a sensible definition of autonomy which considers thought processes, rationality and motivation.

As such, the definition which this thesis will adopt is provided by Gerald Dworkin in his work 'The Theory and Practice of Autonomy'. The definition is as follows:

> "[A]utonomy is conceived of as a second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to

> accept or attempt to change these in light of higher-order preferences and values. By exercising such a capacity, persons define their nature, give meaning and coherence to their lives, and take responsibility for the kind of person they are." – Dworkin, 1988.

Certain elements of this definition are worth highlighting. Firstly, there is a distinction between first-order and second-order preferences or desires. The difference is both chronological and philosophical. A first-order preference is a fleeting desire, a split-second instinctive decision, an impulse, about something that is currently happening (Dworkin, 1988). For our purposes, a first-order preference might be the desire to watch an attractive person dancing, when scrolling through TikTok. A second-order desire is a more stable, reflective, thought-based desire, about a first-order desire (Dworkin, 1988). Importantly, this thought does not have to take the form of a clear philosophical axiom like: 'I shall maintain sovereignty over my will'. As Dworkin notes, if this was the case, only philosophy professors would be autonomous (Dworkin, 1988). Indeed, second-order preferences can be emotionally-driven or vague; their only requirements are criticality and reflectiveness. In the previous example, a second-order preference might be the thought that we wished we were not jealous of, or lustful towards, the person in the TikTok. Dworkin's hierarchy of preferences is more precise than Kant's division of 'reason' vs 'inclination'. It is also appropriate for AI-based systems, which so often manipulate people implicitly, rather than explicitly, playing upon our first-order desires. Dworkin's division could be criticized for being obscure, but not incomprehensible. Although laypeople will not have heard of this division, it is not difficult to understand.

2.3 Autonomy and AI.

Mainstream AI ethics places significant emphasis on autonomy. However, there is an "overall lack of structure in the current discourse" (Prunkl, 2022). Part of this stems from a lack of a unified definition of autonomy. Indeed, the EU HLEG suggests that it is based in protection from "unjustified coercion, deception or manipulation"; the OECD emphasizes "human determination"; and others specify "control" and "decision-making" (Prunkl, 2022). Moreover, autonomy is often imprecisely conceptualized in the field due to cross-over between papers on machine autonomy and human autonomy (Calvo et al., 2020).

Despite a lack of total clarity over the concept, useful explanations of how AI can harm or benefit autonomy can be drawn together from the literature. The benefits are few in number, but profound in significance. Most importantly, AI can reduce the time,

energy, money and effort that individuals spend making small or tedious decisions, thus freeing their minds for larger or more interesting ones (Yang et al., 2018). AI is excellent at performing uniform pre-defined tasks in self-contained 'enveloped' environments (Floridi, 2011); humans can take advantage of this, to eliminate "drudgery" from their lives (Floridi et al., 2018). From Dworkin's perspective, AI automates the satisfaction of basic first-order preferences, ideally in line with second-order desires. Secondly, although AI systems often display information in a manner which has manipulative effects, in many cases its personalization and data-crunching facilitate the provision of relevant and useful information (Bjørlo et al., 2021). This has been repeatedly highlighted in interview studies on AI's benefits to the everyday individual (Sankaran & Markopolous, 2021).

Thirdly, AI can facilitate self-nudging, thus ensuring that an individual's second-order preferences are emphasized in their everyday life (Floridi et al., 2018). An individual might have the second-order preference to exercise every day. A Fitbit is programmed with an operationalization of this preference, and translates it into a first-order entity, by alerting the user to their previous commitment on a continuous basis. Fourthly, some argue that AI systems can expand the options which individuals have, and thus expand their autonomy. This is only partially true. If an AI system provides an individual with a greater range of activities which they are "free to do", then it is their 'positive liberty' which has actually been expanded (Berlin, 1958). However, if additional choices provided by an AI system positively affect an individual's ability to align their first-order desires with their second-order desires, or their critical reflections on said desires, then the individual's autonomy has been enhanced. As such, if an AI allows an individual to apply 100 different snapchat filters to their selfies, their autonomy has likely not been enhanced. However, if an AI auto-translates a foreign news source to provide the individual with life-altering information, or allows them to learn a new language, the individual's autonomy may have been enhanced.

AI can harm autonomy in a wide variety of ways. Most importantly, AI can be manipulative. Indeed, while the algorithms themselves have no 'intention', they are mathematically designed to maximize complex functions, which are in turn based on the implicit or explicit ideas and desires of engineers and managers (Calvo et al., 2020). AI's creators may want the individual to stay on a platform and see more advertisements, or make particular decisions in their daily lives, or think certain things.

As such, AI covertly changes our first-order preferences to be aligned with third parties' desires (Calvo et al., 2020), rather than our second-order preferences. The effects of these changes can range from the seemingly harmless to the profoundly harmful. Individuals may not be concerned about how predictive text shapes how they write online (Varshney, 2020). However, they will be profoundly concerned about forming habits which conflict with their most fundamental values; this often occurs, since a useful mechanism for maximizing engagement is appealing to individuals' base desires, and creating 'sticky traps', whereby a certain type of content is pushed upon an individual until they become trapped in a feedback loop, and seek it out themselves (Milano et al., 2020). Sometimes AI manipulates individuals towards goals which are *thought* to benefit them; for instance, government AI systems often nudge individuals towards official guidance. However, even paternalistic manipulation harms autonomy. It is also worth noting that manipulation can be more harmful when it is biased. Individuals who are manipulated to think in alignment with a feasibly stereotyped version of a particular group of people, cede their first-order desires not just to third parties' desire to make money, but to their social opinions or potentially their discriminatory views. A number of features of AI systems facilitate more effective manipulation. Many AIs misrepresent themselves as neutral machines which provide objective information, which creates a sense of trust (Bjørlo et al., 2021). Moreover, their obvious complexity, the myriad capabilities they can offer, and the frequent rate at which they change, can all overwhelm the user (Friedman, 1996, in Calvo et al., 2020). Furthermore, in some fields, individuals are not in the right 'headspace' to consider algorithmic manipulation due to the intrinsic and immediate distraction that the AI provides (Bjørlo et al., 2021). All of these elements make it more difficult for individuals to observe manipulation accurately.

Alongside direct and intended manipulation, AI systems often lead to "adaptive preference formation" (Prunkl, 2022), whereby an individual's first- and sometimes second-order desires are shaped by the options and categories available, rendering those desires dependent or unauthentic. For instance, a shopping website might allow individuals to filter and sort by price. If its 'low cost' category contains shoes priced between £20 and £50, then an implicit message is being sent, that a frugal shopper *ought* to be spending this much. Thus, the individual's own reflections on how much they should be spending, are adapted by the options with which they are presented.

Since there are many choices which an individual must make when interacting with most AI systems, they will often take the quickest or easiest route, and defer to defaults or not examine their options (Sankaran & Markopolous, 2021).

AI systems can also lead to loss of competence and the establishment of dependency. If individuals regularly defer tasks to AI, they will become worse at those tasks themselves (Prunkl, 2022). For instance, if individuals base their viewing preferences on Netflix's recommendations, they will become worse at finding good films on their own. Knowledge of which directors and actors are worth watching, which film reviewers are worth listening to, etc., will atrophy. As such, dependency is established, which reduces our ability leave our current system, or function when it fails (Prunkl, 2022). In other words, our ability to change or reliably perform certain actions declines; thus, our autonomy is reduced.

The field of AI ethics does not provide useful operationalizations relating to autonomy for everyday individuals; however, certain scholars make vital observations, which facilitate the creation of said operationalizations. Calvo et al. suggest the use of a framework from the Human-Computer Interactions literature, called 'METUX', which divides experience of technology into six spheres: 'adoption', 'interference', 'task', 'behaviour', 'life', 'society' (Calvo et al., 2022). Certain stages associated with each sphere might leave the individual more vulnerable to autonomy harms than others. Bjørlo et al. suggest that AI ought to promote 'complementarity', i.e., the achievement of tasks which were already intended for completion, made easier via AI; as such, all "judgement" is left to the person, with the AI as a simple machine (Bjørlo et al., 2021). Finally, Bjørlo et al. observe that transparency enhances reason or judgement (Bjørlo et al., 2021). Judgement in turn, in Dworkin's language, facilitates the formation of second-order preferences, and the evaluation of first-order preferences, thus strengthening autonomy. Although transparency is important in its own right, this establishes symbiosis between the two values.

2.4 Transparency from Below: Possessing Knowledge and Understanding

The field of epistemology is vast with limited consensus arising from centuries of complex arguments. As such, while it was possible for this thesis to consider various ideas of autonomy before selecting a particular vision, all conceptions of knowledge and

understanding cannot be explained here. Instead, this thesis will carve out a path through the literature and justify taking it.

Although there is considerable debate on the ultimate value of knowledge and understanding it is generally presupposed that both entities are intrinsically desirable. The possession of knowledge holds instrumental pragmatic value for the knower (Pritchard, 2009). Indeed, being 'right' about something, for instance the path to a particular location, either via knowledge or mere 'true belief', insulates an individual from the likelihood of failure and maximizes one's chances of success in achieving a pre-established goal (Pritchard, 2009). Along a similar vein knowledge of one's options facilitates informed choice, knowledge of oneself facilitates welfare, and knowledge of the previously unknown can soothe anxieties (Turri, Alfano & Greco, 2021). Knowledge and understanding also hold value as the aims of human activities: acquiring knowledge satisfies our innate curiosity and constitutes success in any act of inquiry (Kvanvig, 2003). Moreover, reliabilist theorists have suggested that knowledge's value lies in its longevity. Indeed, since many situations repeat themselves throughout our lives we can accumulate knowledge of how to deal with them best and repeatedly reuse it (Olsson, 2011). It is further worth noting that alternatives to states of knowledge include indifference, ignorance, and falsehood all of which are often seen to be detrimental to an individual's life.

Although there are many alternatives this thesis will use a baseline definition of knowledge from Linda Zagzebski's 'Virtues of the Mind'. Here it is argued that:

> "Knowledge is a state of cognitive contact with reality arising out of acts of intellectual virtue." – Zagzebski, 1996

Several elements of this definition are worth highlighting, especially due to their contrast with a traditional definition of knowledge as "justified true belief" (Audi, 2011). Firstly, this conception of knowledge excludes beliefs which happen to be true but are not examined or arrived at by deliberate and "non-accidental" means (Zagzebski, 1996). As such, if an individual colloquially 'knew' the result of the election before it had occurred, but ended up being correct, this would not constitute knowledge. Similarly, if a person 'knew the way' to a location, and ended up successfully guessing a route, this would also not constitute knowledge. With this said, this thesis' interpretation of 'non-accidental' will be forgiving. As such, learning about

an AI system's operations by half-listening to an auto-played YouTube video, will not be considered 'accidental'. In this circumstance, it will be assumed that the act of paying some attention to a source constitutes a bare minimum of 'intellectual virtue'. This feature of Zagzebski's definition is particularly useful for the purposes of this thesis, since so many people 'think they know' about AI or about the entities which create important algorithms in their lives, without actually possessing knowledge about them. Secondly, it should be noted that this definition does not tie knowledge to 'the truth' but instead to 'reality'. This is important since in many AI ethics cases there is no singular 'truth' which can be uncovered but simply a reality which we need to try and get as close to as possible (Ananny & Crawford, 2018). A classic example is that of the 'black box' AI system in which a complex machine learning or deep learning algorithm changes and makes decisions which cannot even be explained or understood by its creators (Mittelstadt et al., 2016). Here, the reality of the system is of the utmost importance and finding a singular truth is impossible.

Zagzebski's definition outlines what has occurred when an individual possesses a piece of knowledge. However, it is also important to understand where knowledge comes from and how knowledge is structured. Robert Audi identifies several sources of knowledge. Most importantly, there is testimony whereby an individual may believe 'p' because an external source 'S' has informed them of it (Audi, 2011). In the case of AI ethics from below 'S' may be a peer-reviewed article, a newspaper, a podcaster etc. The credence which we offer to 'S' should depend upon whether the information they provide seems prima facie reasonable and upon the source's track record, reliability, verifiability and process (Steup & Ram, 2020). An individual may also gain knowledge from reason or the use of logic to come to certain conclusions (Audi, 2011). They may further derive knowledge from their immediate perceptions of the world around them, their memories, and introspection of their own minds (Audi, 2011). Importantly, none of these sources are mutually exclusive and all of them may be combined to provide a better justification for a piece of information or a belief comprising 'knowledge'. This will be important in this thesis' consideration of the gathering of information by the individual.

Once it is gathered knowledge needs to be mentally stored. Two theories of the structure of knowledge are worth noting. The first is foundationalism which conceptualizes knowledge as basic and non-basic. Basic knowledge, which receives no justification

from any other piece of knowledge, acts as a foundation for a superstructure of non-basic knowledge, which receives justification from other pieces of knowledge (BonJour, 2017). The second is coherentism which conceptualises the pieces of knowledge which an individual possesses as a web with each piece holding up another (Steup & Ram, 2020). These theories will be useful for conceptualising individuals' ideas about AI systems in the chapter 4.

Finally, it is worth noting the difference between knowledge and understanding. This thesis will adapt its definition from Stephen Grimm's and suggest that understanding is the "mental" ability to consider "how the aspects of a system depend upon one another" (Grimm, 2011). This definition eliminates non-explanatory understanding, i.e., 'understanding that' something is the case, which ought to be equated with belief (Baumberger et al., 2017). There are several elements of 'understanding' which are worth noting. Firstly, it is possible for an epistemic agent to have a false understanding of an entity. To avoid this, it is vital that understanding is based on reality or simply upon what this thesis defines as knowledge. Secondly, understanding something is a more profound state than knowing something. To understand an entity an epistemic agent must intellectually "grasp" its workings and "see" inside its operations (Grimm, 2006). As such, understanding is hierarchically more valuable than knowledge (Kvanvig, 2017). Finally, unlike knowledge which one possesses or doesn't, understanding "admits of degrees" (Pritchard, Turri & Carter, 2018). Moreover, incremental increases in one's understanding are seen as beneficial. This distinction between knowledge and understanding is important. Depending on contextual factors this thesis' framework may encourage either accumulating knowledge or advancing understanding.

2.5 Knowledge and Understanding Through AI Transparency and Explanation

In alignment with the field of AI ethics as a whole, scholars of knowledge and understanding of AI view these concepts from a 'top-down' perspective. As such, institutional transparency and explainability are key foci of the literature. Although there is disagreement over the definitions of these practices and their interrelation it is worth picking out some useful strands from the discourse.

Transparency should be understood as the provision of information about the inner "workings and performance" of a system to external agents (Diakopolous, 2020).

Meanwhile, explainability and associated terms such as "intelligibility, comprehensibility, understandability" and "foreseeability", involve deeper guided comprehension of these inner workings, generally based on the description of causal mechanisms (Wischmeyer, 2020). As such, we may say that as knowledge is to understanding, transparency is to explainability. Importantly, when AI systems are opaque to the public individuals can suffer epistemic harms such as ignorance and undue indifference and are prevented from the epistemic successes detailed in the previous section. Moreover, being informed is often a necessary factor in the protection of other values such as autonomy, privacy, welfare, and justice (Ananny & Crawford, 2018).

Despite near-universal endorsement of these values, organizations which create or use AI are often reticent to practice them. From a business perspective there is significant risk associated with explaining the operation of a carefully developed algorithm to one's competitors (Wachter, 2018). There are also monetary costs associated with trying to understand and communicate every choice made in an AI system's design (Diakopolous, 2020). Moreover, organizations have a natural incentive to hide aspects of algorithms which could be seen as unethical (Zuboff, 2015). As such, algorithmic opacity is widespread.

While the benefits of transparency and explainability and the harms of opacity around AI systems are relatively self-explanatory, there are several other elements of the traditional AI ethics literature which should certainly be noted for the construction of AI ethics 'from below'.

Firstly, an important task of the existing literature lies in deciding what should and what should not be transparent. It is clear that conceptions of transparency based on laying bare every facet of an organization's products and functions to the public are unrealistic (Theodorou, 2019). Instead, transparency must be conceptualized as a sliding scale based on selective disclosure or research (Diakopolous, 2020). Two perspectives on the extent of transparency should be considered for this thesis.

In its most extensive state transparency should involve the display of several key features of AI systems to the public. Most basically, a transparent organization would disclose whether an algorithm is or is not functioning in any given application (Ausloos et al., 2018). Next, it would detail all human involvement in an AI system's design and

operation, from the decisions made by engineers to the assumptions and intentions of the project managers to the funding sources gathered by executives (Diakopolous, 2020). Alongside this, the "organizational goal" of the algorithm, i.e., the end which it attempts to maximize or achieve, would be explained (Diakopolous, 2020). Then, the type of data used by a system, as well as quality assurance standards around its "accuracy, completeness, timeliness and update frequency" would be disclosed (Diakopolous, 2020). Finally, the inner workings of the AI system itself, including but not limited to the outcomes it produces, the features it uses, the type of inferences it makes, and the type of model it is, would be explained as well (Wischmeyer, 2020).

Although this acts as an effective 'upper bound' for transparency and explainability it is worth noting that there are possible harms associated with such thorough information disclosure. Indeed, explaining everything about an AI system can overwhelm and confuse the reader (Ananny & Crawford, 2018). Even worse, extensive disclosures or transparency buzzwords like 'open-sourcing' can lead individuals to fall back on 'heuristic cues' such as the assumption that 'if a document is long it is likely to be accurate' and place their trust in entities without any knowledge or understanding of how they work (Liu, 2020). In other words, transparency can lull individuals into a false sense of security. As such, other voices within the field argue for more concentrated, concise disclosures or explanations. Various scholars, alongside landmark regulations such as GDPR, suggest that the 'logic' of the algorithm is most important: i.e., its existence, its aim and how that aim is fulfilled (Preece, 2018; Ausloos et al., 2018; Wischmeyer, 2020). Others suggest eliminating explanations of processes and focusing entirely on outcomes (Diakopolous, 2020). Although these more concise explanations sacrifice "fidelity" to the reality of the situation to "interpretability", it is generally seen that the latter is more important for the layperson (Wischmeyer, 2020). For the purpose of this thesis, it is useful to take both the extensive versions of transparency, and the more limited ones, into account. Individuals may wish to start with concise explanations before seeking more detail.

Alongside discussions of what should be transparent, scholars emphasize that the provenance, form, and nature of the information disclosures themselves should be examined. Information that is leaked by a whistle-blower will be unfiltered; information proactively sought out by an individual via the exertion of their rights will often be accurate but unfocused; information which is provided for compliance may provide the

bare minimum of detail possible (Diakopolous, 2020). These observations will be useful for the frameworks' recommendations regarding research practices.

To facilitate the creation of AI ethics 'from below', it is also worth outlining concerns within the literature regarding the feasibility of instigating effective transparency. For instance, there are a number of technical obstructions to this end. Algorithms are often dynamic entities, which are changed when new training data is used, or when new updates arise; as such, understanding an algorithm at one time, might not equate to understanding it a month later (Kemper & Kolkman, 2019; Diakopolous, 2020). Moreover, algorithms are often irreducibly complex. Webs of different types of models, interlaced with hundreds of different human agents with feasibly different motivations and assumptions, can be very difficult to understand (Wischmeyer, 2020). In some cases, it may even be impossible to tie a given input to a given output (Wischmeyer, 2020).

A further barrier to the instigation of effective transparency is the potential for transparency and explainability to have unethical effects. Transparency can facilitate the gaming of AI systems, as individuals can either act deceptively or manipulate the services themselves to gain beneficial results (Diakopolous, 2020). While this itself is not unethical it becomes so if others are disadvantaged in the process or if the agent causes themselves harm by altering something they do not truly understand. Transparency around data can also lead to invasions of privacy (Diakopolous, 2020). It is important to anonymize information when it is scrutinized.

Despite the hindrances to their efficacy acknowledged by the field, the AI ethics literature does offer some useful mechanisms for operationalizing transparency and explainability. Firstly, there are a selection of methodologies which involve a guided expert or designated AI system working individuals through information relating to a certain algorithm. "Hypothetical" or "counterfactual" dashboards are good examples of such mechanisms (Wischmeyer, 2020). These resources show users how potential changes to their behaviour or an algorithm's function alters outcomes or decisions (Wischmeyer, 2020). Creating such resources is time intensive, and their focus on highly specific information makes their contributions non-transversal. Secondly, technical disclosures have been recommended. For instance, publishing source code, providing information on training data, disclosing biases ahead of time and showing

individual activity logs, may all aid the aims of transparency (Beaudouin et al., 2020). The wider comprehensibility of this sort of information is questionable. Thirdly, although they are often not mentioned within the corpus on transparency, ‘ethical frameworks’ are often used to explain AI ethics to engineers, legislators and c-suites. These guides provide conceptualizations and operationalisations for users to employ, and work them through multiple stages of change, with the end-goal of more ethical behaviour in regard to AI systems (Morley et al., 2021). It is this mechanism, which this thesis adapts for the individual.

3. Methodology

3.1 "Intended Knowledge"

This study utilized semi-structured interviews with AI ethicists[1]. Alongside a thorough literature review and analysis of key philosophical texts, these interviews strongly guided the creation of this thesis' ethical frameworks. Each interview was split into two sections.

The first section was designed to derive information about AI ethicists' behaviours around AI. This section was composed of four questions (see table 1). It was believed that AI ethicists protect and enhance their autonomy around AI, and accrue information about it, more effectively than average individuals. As such, they were able to provide 'model operationalisations' for laypeople to imitate. These were habits, behaviours and ideas, generated by experts, which verifiably worked with the practicalities of everyday life. Additionally, interviews revealed differences between individuals' perspectives and values. Acknowledgement of areas of heterogeneity facilitated the creation of frameworks which should work for everyone.

Table 1 – First Section Interview Questions

Question 1	If you were advising a close friend, who did not know about these sorts of things, about changing their behaviour around AI systems, what might you say?
Question 2	Are there any specific behaviours which you, or other people you know in the field, have changed since learning about the effects of AI?
Question 3	Is there any particular service which you, or other people you know in the field, are particularly wary of?
Question 4	Is there any piece of information in the field which piqued your interest at an early stage and made you want to learn more?

Following other studies which have used interviews to "co-produce" knowledge regarding sociotechnical aspects of algorithmic technologies (Jia et al, 2012; Edwards &

[1] CUREC approval (reference SSH_OII_CIA_22_047) was received for this project.

Holland, 2013; Hand, 2014), the second section of each interview was designed as a site of discourse. This section was composed of semi-structured questions, tailored to each interviewee. In a pilot study for this thesis, I discovered that tailoring schedules based on interviewees' prior publications and research interests prevented the regurgitation of standard AI ethics in interviews and facilitated focused in-depth conversations (Candidate 1061912, 2022). Tailoring schedules also increased interviewees' attentiveness, immersion, engagement and insightfulness as it signalled my own expertise, as well as openness to and interest in their ideas (Hand, 2014; Howlett, 2021). With this said, discourse remained reflexive and open, as was appropriate given the nuanced and diverse nature of the subject matter at hand (Kazmer & Xie, 2008).

Discussions in the second sections of each interview facilitated the provision of commentaries and criticism of my ideas as they developed; allowed for tests and assessments of potential elements of the framework to be provided in real time; and introduced new ideas about AI from disparate fields such as psychology, design, computer science, epistemology, business studies and law. Finally, they facilitated dialogue involving honest opinions about the discipline of AI ethics: thoughts which are kept out of finished papers and publications, which account for practical conditions and set aside the aspirations of the field.

3.2 Sampling

This study employed a 'theoretical sampling' technique. As such, participants were selected based on "their (expected) level of new insights" (Flick, 2009). Interviewees' previous publications, projects or jobs had to be explicitly and strongly related to AI ethics. Interviewees could, but needn't be, directly engaged in research into autonomy, transparency or explainability. To avoid conflicts of interest, no members of faculty at the Oxford Internet Institute were included.

Potential participants were identified either through prior contact; or via searches for "AI Ethics", "AI Ethics Autonomy" and "AI Ethics Transparency" on Google Scholar, IEEE Xplore and the ACM Conference database; or through examination of webpages of relevant institutions such as the Leverhulme Centre for the Future of Intelligence. All recruitment emails were personalised, to show my knowledge of potential interviewees' work and my own expertise in the field (Harvey, 2011). Over 30 AI ethicists were contacted for this study. A full list of those interviewed can be found in table 2.

Table 2 – Interviewees and their Affiliations
Ben Gilburt, AI Ethics Lead at Sopra Steria, Founder of AI Ethics London, MSc Student in the Social Science of the Internet at the Oxford Internet Institute of the University of Oxford
Lena Vatne Bjørlo, PhD Candidate in Digital Marketing at the Department of International Business of the Norwegian University of Science and Technology
Elena Falco, PhD Candidate in Science and Technology Studies at University College London, Research Assistant at the Leverhulme Centre for the Future of Intelligence
Florian Richter, Research Fellow at Technische Hochschule Ingolstadt
Fabrice Muhlenbach, Associate Professor in Computer Science at the Herbert Curien Laboratory of Jean Monnet University
Maya Sherman, Former Research Assistant at Interdisciplinary Center Herzliya, MSc Student in the Social Science of the Internet at the Oxford Internet Institute of the University of Oxford
John Zerilli, Leverhulme Fellow at the University of Oxford, Associate Fellow of the Leverhulme Centre for the Future of Intelligence at the University of Cambridge
Elizabeth Seger, PhD Candidate in History and Philosophy of Science at the University of Cambridge, Research Assistant at the Leverhulme Centre for the Future of Intelligence
Markus Langer, Post-Doctoral Researcher in Psychology at Saarland University

3.3 Execution & Analysis

Interviews were held either face-to-face or via Zoom and lasted between 40 and 60 minutes each. A synchronous 'live' approach was necessary since my comprehension of interviewees' explanations of their values, beliefs, behaviours and habits required close attention to tone and non-verbal cues (Hewson, 2014). This approach also facilitated 'steering' from one topic to another, which was necessary given the diversity of topics covered (Brinkman & Kvale, 1996). Moreover, in this format several academics reverted into 'supervisor' roles. This facilitated thorough and nuanced discussions and was accompanied by useful behavioural norms such as pausing to think, looking up technical concepts, and sending links to useful resources.

Field notes were taken by hand and typed up immediately after each interview. Every interview was fully or partially transcribed. After all interviews had concluded, the first sections of each interview were coded using a combination of descriptive, theoretically-driven a priori codes and data-driven a posteriori codes (Saldana, 2013). This coding process primarily sought to find frequencies, similarities and differences (Hatch, 2002). The second section of each interview was coded with a posteriori codes only, and did not seek to find patterns, themes or commonalities, but to generate summaries of unique viewpoints (Patton, 2002).

Since the aim of this thesis was not to examine and discuss how a particular set of AI ethicists behave and think, findings are not presented in a 'results' section. Key insights and findings from the coding process are "interwoven" into the following chapters and triangulated with key philosophical perspectives and contemporary AI ethics (Brinkman & Kvale, 2015).

3.4 Methodological Limitations

There are a number of limitations to this methodology.

Firstly, since this study allowed contacted interviewees to self-select into the study unknown biases may have been introduced (Lieberson, 1987). It is possible that scholars who believed in the potential of the framework were more inclined to participate. It is also possible that certain types of AI ethicists were more disposed to collaboration than others; this would explain why, although roughly equal numbers of each were contacted, only one scholar with an explicit interest in autonomy agreed to be interviewed, while four with foci on transparency and explainability did so. Potential imbalances were considered when constructing the next chapter's frameworks.

Secondly, the sample collected may introduce specification error to this study. This occurs when there is a difference between the "concept" "needed to address a research question" and the "concept implied by the data item" (Amaya et al., 2020). This thesis aims to create ethical frameworks for *all* people. However, those interviewed may not be representative of *all* people. Although the interviewer was a person of colour, no interviewees were. All interviewees were currently employed or being educated in Europe and lived in free democracies. Although it would be wrong to speculate about the socio-economic status of interviewees, it is possible that some backgrounds were un- or under-represented. Ages ranged from 20s to 60s, and the gender balance of the

interviews was 5:4 male to female; however, some ages and gender identities were still unrepresented. The implications of these differences were considered in writing the next chapter, in the interests of maintaining inclusivity.

It is worth noting that the nature of this study meant that a broad sampling technique was required. The author of this thesis was an MSc student rather than someone of higher academic status, contacting elites at prominent institutions to take part in a study of no tangible benefit to them, at a busy time of the academic year. Responses could only be prompted through intensively researched emails or prior contact. Even a slight restriction on the sampling technique, such as the use of quota sampling, would have led to fewer interviews of a lower quality. With this said, its flaws should nonetheless be acknowledged.

A final limitation of this methodology was the inextricability of my own biases from the interview processes (Van Haitsma, 2009). Although I based my questions in interviewees' own work, the AI ethics literature, and well-respected philosophical arguments, my readings of all of these entities may have been skewed.

With all of this in mind, it is worth pointing out that this thesis does not claim the generalizability of its results either regarding AI ethicists' views and behaviours or the lessons individuals should derive from them (Blank, 2017). It simply presents a good set of options based on one set of interviews. It should also be reiterated that interviews form just one guiding entity for the following frameworks. Philosophical analysis and a review of the AI ethics literature were equally important in their creation.

4. AI Ethics Frameworks for the Individual

This chapter presents two ethical frameworks designed to support laypeople in their interactions with AI. The first is centred around knowledge and understanding. The second is centred around autonomy. After considering key challenges to AI ethics operationalizations and how to overcome them, each framework is displayed as a flowchart, explained and justified, and presented in full.

4.1 Challenges to Constructing AI Ethics Frameworks

Following Morley et al. (2021), I have identified eleven challenges which past AI ethics frameworks, principles documents, education schemes and laws have faced. These challenges are detailed in table 3 below. Understanding the hindrances and pitfalls associated with previous operationalizations is key to creating effective frameworks.

Table 3 – Challenges to Constructing AI Ethics Frameworks

Challenge	Explanation
Ignorance & Apathy	Stakeholders may not possess baseline knowledge about AI or AI ethics. Furthermore, they may be unwilling to learn (Wischmeyer, 2020; Liu, 2021).
Enforceability	It is not mandatory for stakeholders to follow or employ AI ethics principles or guidelines (Theodorou & Dignum, 2021).
Comprehensibility	Stakeholders lack clear and unified definitions of core values, concepts and operationalisations (Mittelstadt et al., 2016; Jobin et al., 2019).
Non-prescriptiveness	AI ethics resources may describe problems, without prescribing solutions (Hagendorff, 2020; Morley et al, 2021).

Lag	AI ethics resources often have a "long lead time". This can lead to extensive periods of non-guidance at best, and instant obsolescence at worst (Morley et al, 2021).
Excessive Flexibility/Rigidity	AI ethics guidance may be excessively fluid, allowing individuals to apply them as post hoc justifications for any action taken. They may also be excessively rigid, limiting their transversal applicability (Morley et al., 2021).
One-Off Solutions	AI ethics resources may be presented as one-off solutions (Morley et al, 2021).
Bad Faith	AI ethics resources may be written in bad faith (McMillan & Brown, 2019; Schiff et al., 2020). "Ethics shopping" or "ethics washing" may be occurring (Floridi & Cowls, 2019).
Unethical Ethics	AI ethics resources may be unethical from some perspectives (Roberts et al., 2019).
Exclusion of Key Stakeholders	Certain stakeholders' interests may not be considered by AI ethics resources (Sherman, 2022; Muhlenbach, 2022 (Interviews)).
Value Trade-Offs	AI ethics resources may not consider value trade-offs in sufficient depth (Falco, 2022 (Interview)).

Engagement with laypeople presents three additional challenges. These are reported in table 4 below.

Table 4 – Further Challenges

Challenge	Explanation
Limited Cognitive Capacity	Individuals may not have the cognitive capacity to consider complex ideas around AI or AI ethics (Falco, 2022 (Interview)).
Limited Time	Individuals may have limited time to employ the framework (Kemper & Kolkman, 2018).
Potential Paranoia or Luddism	Individuals may become paranoid about AI-producing organizations. They may also abandon certain technologies or services without proper consideration (Zerilli, 2022 (Interview)).

The proposed frameworks must account for these challenges. Individuals' engagement with the frameworks must be incentivised via descriptions of the benefits of their use. Relevant terms should be defined clearly and jargonistic technical language should be avoided. The frameworks must be composed of transversal principles, step-by-step processes and questions. Operationalizations must be simple to account for users' limited time and cognitive capacity. Rigid rules and imposed hierarchies must be evaded. As such, individuals should be able to use the framework in relation to any service, at any time, in most contexts. With this said, the framework cannot be excessively permissive; the results of its use should be similar for most users. Moreover, at pertinent junctures the framework should widen its perspective, allowing for the consideration of other values and/or stakeholders. Finally, daunting terms should be avoided where possible. Interviewee Maya Sherman, a scholar of intersectionality and AI, pointed out that laypeople may associate the word "ethics" with repellently complex academic writing. As such, "values" or "effective behaviours" will be discussed instead. Moreover, as Markus Langer, a psychologist who has worked on receptivity to different terms around AI, pointed out in his interview: the term "AI" can be useful for driving engagement, but "algorithms" are easier for people to deal with conceptually. This thesis' definition facilitates the interchangeability of the terms.

4.2 Knowledge & Understanding

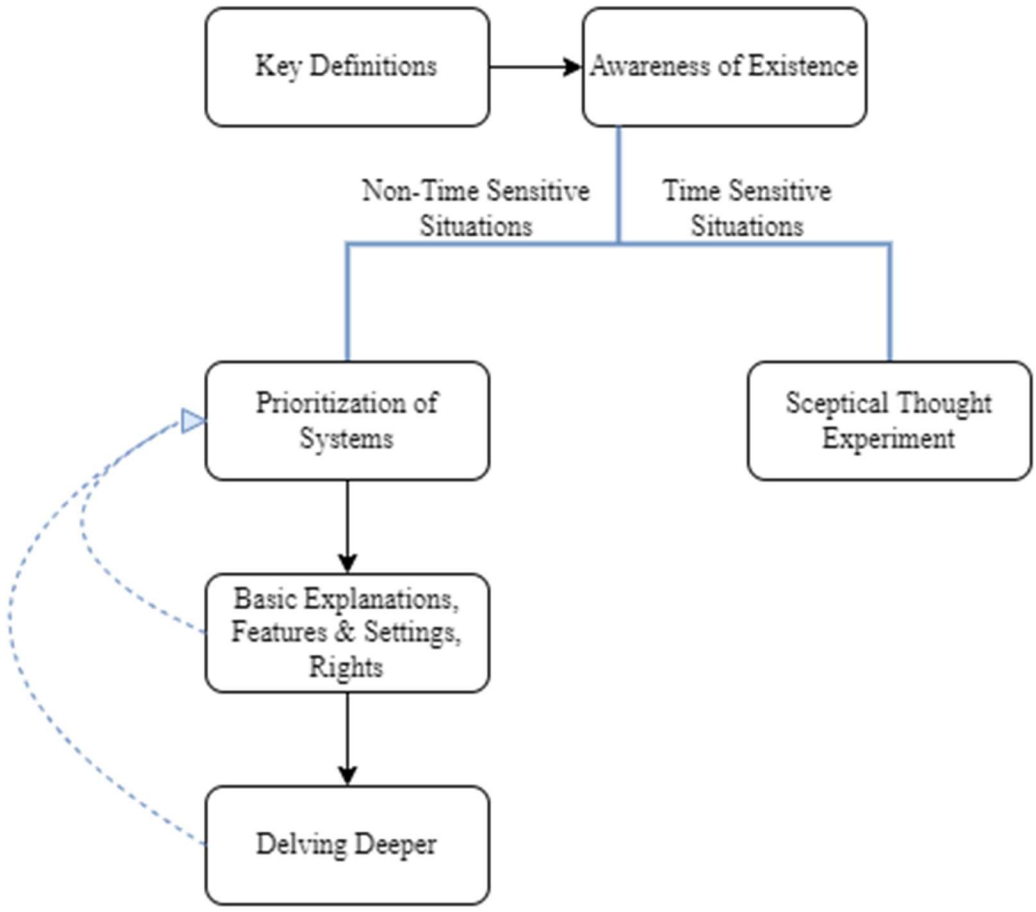

Figure 1: A Framework for Knowledge and Understanding (Flowchart)

Figure 1 displays a flowchart representation of the framework for gaining knowledge and understanding of AI systems. Each stage of this flowchart is explained in the following section.

4.2.1 Key Definitions & Foundational Knowledge

From a foundationalist perspective it is important that the "basic knowledge" upon which individuals' ideas about AI are constructed is grounded in reputable AI ethics (Bonjour, 2017). As multiple interviewees raised, it would be idealistic to assume that individuals do not have inaccurate preconceptions or significant blind-spots regarding AI.

Firstly, individuals must possess a clear, practical, socio-technical definition of AI. Here Floridi & Taddeo's (2018) definition is offered. Next, harmful narratives must be

dismissed (Shin, 2021). Individuals may think of AI as a future, rather than present, concern (Fast & Horvitz, 2017). They may associate AI with science fiction and believe that it must have an active "persona" (Hermann, 2020); they may be concerned with "losing control", AI sentience, or the singularity (Recciha, 2020). As interviewee Fabrice Muhlenbach pointed out, many professionals may be concerned with AI *only* in relation to 'taking their jobs'. All of these ideas must be ruled out in the framework. In conjunction, individuals must understand the ubiquity and power of AI in their lives. Description of AI's everyday effects will both direct their interest to key topics and incentivise them to use the framework.

4.2.2 Awareness of Operation

In his interview, Florian Richter raised a quotation by ubiquitous computing scholar Mark Weiser: "the most profound technologies are those that disappear", "they weave themselves into the fabric of everyday life until they are indistinguishable from it" (Weiser, 1999). Alongside two other interviewees he pointed out that laypeople do not conceive themselves to be "using AI" when they adopt or employ AI-based systems. In order to consider AI's effects transversally, select particular services to examine, and indeed employ the rest of the framework, individuals must first understand which services use AI and which features signal that an AI is in use (Ausloos et al., 2018). As such, the framework provides individuals with a list of product types which are likely to use AI.

4.2.3 Non-Time Sensitive Learning: Prioritization

Individuals' incentives to learn about AI are increased if it has a direct connection to their lives. As such, this framework advocates a service-by-service approach. This method acknowledges that individuals' experience of AI systems cannot be entirely disconnected from the products in which they are embedded (Kemper & Kolkman, 2018). It also facilitates incremental deepening of individuals' understanding (Grimm, 2011). Indeed, it is anticipated that individuals who iterate back and forth between different services will accrue generalizable knowledge about AI models in the long-run. As such, later iterations should be easier than earlier ones. This approach also circumvents issues regarding the difficulty of learning about AI purely via general theory about different types of models. Moreover, while facilitating the transversality of the framework, a system-by-system strategy allows individuals to consider the intricacies of particular AIs in heterogenous fields (Eitel-Porter, 2021).

Individuals must prioritize systems to research based on actual and anticipated impact (Beaudouin et al., 2020). This was a commonly advocated practice amongst interviewees. In doing so users should consider the time they regularly spend with the AI system (Calvo et al., 2020); its emotional impact (Jobin et al., 2019); its impact on decision-making (Wischmeyer, 2020), beliefs and motivations (Prunkl, 2022); the reputation of the organization behind it (Schiff et al., 2020); and the importance of the activity it undertakes to their lives and the health of wider society (Yang et al., 2020).

4.2.4 Non-Time Sensitive Learning: Initial Research, Gaining Knowledge

This thesis advocates a targeted research-based approach to gaining knowledge and understanding. As such, individuals are asked to use lay-accessible resources to compile information on a set number of sub-topics. There are several justifications for this approach. Firstly, it is important to avoid a "logic of accumulation" in relation to knowledge (Ananny & Crawford, 2018). More information is not necessarily better information. Moreover, as interviewees Ben Gilburt and John Zerilli raised, information overload at an early stage can lead individuals to become overwhelmed at best or paranoid at worst. Providing research targets prevents this from occurring. Secondly, individuals must not be allowed to gather information in an unfocused manner. This could lead to distorted or incomplete perspectives on AI, with their particular interests gaining exaggerated importance in their minds. Guiding individuals to research key topics in academic AI ethics will ensure that their views are well-calibrated. Thirdly, an 'at home' research-based approach is applicable for almost any type of AI system and requires little-to-no expertise on the part of the user. The same cannot be said for more complex 'explainability' solutions advocated elsewhere in the field (Wischmeyer, 2020; Hagendorff, 2020).

With this said, there are various dangers associated with home research, as interviewee Elizabeth Seger pointed out. Misinformation, ethics-as-marketing from AI-producing organizations, unfounded opinion, and business- or profit-oriented explanations can all lead to epistemic harms. In order to mitigate potential risks, individuals must be advised to follow epistemically secure practices. Individuals are asked to check all testimony online against their own sense of reason and their experiences, memories and thoughts of using AI systems (Audi, 2011). They are asked to consider the reputation and track-record of the publications and authors they read (Seger et al., 2020); to avoid fringe opinions from "insular online communities" (Seger et al., 2020); to cross-reference all

information between different sources (Seger et al., 2020); to consider the context in which different pieces were written (Diakopolous, 2020); and to compile information into a written document to ensure that it remains structured and to facilitate re-use (Kemper & Kolkman, 2019).

Initially, the framework suggests that individuals focus on three aspects of particular AI systems: basic explanations, features & settings, and rights.

4.2.4.1 Basic Explanations

When an AI system is produced by a well-known private organization (Google, Facebook, Twitter, etc.) individuals should first learn about the political economy of that entity. Three interviewees pointed out that the economic workings of major companies will be more intelligible to laypeople than technical accounts of AI. As such, this serves as an effective entry point. Explanations of these issues are plentiful online.

After this, or when an AI system is not produced by a well-known private organization, individuals should research the logic of the algorithm in question. This comprises its existence, its aim, and how that aim is fulfilled (Preece, 2018). For less well-known AI systems it may also be useful to research the type of algorithm rather than a specific service. As such, users should search for "shopping recommendation algorithm" rather than "TopShop algorithm".

Individuals researching the logic of algorithms are hardly asked to examine source code. Instead, they are directed to read the explanations or opinions of journalists, tech experts and academics, for whom and from whom the logic of algorithms ought to be comprehensible (Cath, 2018; Wischmeyer, 2020). Select websites are recommended. These sites are not perfect: they fail to reach the standards of academic AI ethics discourse and may even contain inaccuracies. However, for providing basic knowledge, which can be triangulated and verified between sources, they are useful resources. Moreover, as detailed in Chapter 2's discussion of Zagzebski, when dealing with a subject as complex as AI 'getting closer to the reality', rather than having 'the truth', is a reasonable aim.

4.2.4.2 Features & Settings

Interviewee Fabrice Muhlenbach pointed out that many AI-producing organizations allow individuals to modify how their systems are experienced. For instance, Twitter allows users to change their feeds from 'recommended' to 'latest' tweets. Amazon

allows users to signal that certain purchases should be disregarded for future recommendations. Instagram allows users to turn off 'like-counts'. These are positive *opt-in only* changes. As such, it is imperative that individuals know the full extent of AI-based services' features, settings and options. Moreover, making these changes allows individuals to see, albeit shallowly, into the AI's inner workings.

4.2.4.3 Rights

Alongside well-acknowledged general benefits to knowing one's digital rights, there are two advantages to this practice from a knowledge-seeking perspective. Firstly, resources found on government or regulators' websites about individual rights in relation to the internet can guide individuals towards key issues they should be concerned about, such as privacy, online safety, bias, sustainability and questionable use of data (Diakopolous, 2020). Secondly, knowledge of rights is essential for their exertion (Ausloos et al., 2018). Exertion of key rights, such as the right to access, can provide individuals with first-hand compelling evidence of the practices of AI-producing organizations (Wischmeyer, 2020).

4.2.5 Non-Time Sensitive Learning: A Deeper Understanding

Upon this knowledge individuals may construct a deeper understanding. This phase of the framework provides individuals with various targeted topics which will help them to "grasp" the inner workings of AI systems but leaves them to explore them freely (Grimm, 2011). Most of the recommended subjects are derived from the work of Nicholas Diakopolous, as detailed in Chapter 2. Beyond this, individuals will also be directed to discourses around AI. In particular, technology policy debates and AI regulation will be raised as potential areas of interest. Keeping these elements out of the initial stage of the framework ensures that individuals do not feel burdened with a call to activism.

4.2.6 Time-Sensitive Situations: A Thought Experiment

There will be scenarios in which individuals cannot learn about an AI system before using it. Here the framework proposes an epistemologically conservative, sceptical thought-experiment for individuals to engage with. Multiple interviewees noted the value of simulation or imagination for critical reflection. Moreover, there is precedent for thought experiments being effective in matters of AI transparency (see Theodorou, 2019). As such, individuals are asked to consider what goals the organization which

created the system might have besides providing the service; whether they object to said goals; and whether said goals could lead to harms. Next they should consider whether their objections are strong enough to justify non-use of the system, and if not, whether they can change their behaviour to minimize ill effects. If an AI system has unethical aims or effects the benefits of employing this thought-experiment are self-evident. If it does not, individuals are unlikely to inflict harm upon themselves by conducting a quick thought-experiment. Moreover, since it involves clear guesswork, risks of being misinformed or paranoid as a result of this thought-experiment are not significant. Individuals are told that this experiment should not replace research and that their thoughts about harms and goals can never be exhaustive.

4.2.7 Disclaimer

Finally, the framework issues a disclaimer. Individuals are reminded not to use knowledge unethically. In particular, they are dissuaded from "gaming" systems (Diakopolous, 2020). This can be ethical, so long as others do not suffer from it. For instance, manipulating an algorithm so that you get better recommendations while online shopping is reasonable. However, manipulating an algorithm in order to exaggerate your suitability for a job at the expense of other candidates would not be. Individuals are also informed that their knowledge of AI cannot cover all systems and that due to the dynamism of the field some of their knowledge may quickly become obsolete (Ananny & Crawford, 2018). Accordingly, they are strongly warned against complacency (Heald, 2006).

4.2.8 The Framework for Gaining Knowledge and Understanding Relating to AI

The box below presents the full framework for gaining knowledge and understanding relating to AI, as it would be distributed.

A Framework for Gaining Knowledge and Understanding Relating to AI

Stage 1: Key Definitions and Background

Artificial intelligence (AI) can be defined as:

"A growing resource of interactive, autonomous, self-learning agency, which enables [computers and alike] to perform tasks that otherwise would require human intelligence to be executed successfully" (Taddeo & Floridi, 2018).

AI is a present real-world concern. Science-fiction perceptions of AI should be abandoned.

AI systems are not sentient and do not need to have personas or personalities. There is no risk of 'losing control' to the machines through the singularity.

In reality, AIs are pervasive and powerful technologies which mediate everyday life. AI systems affect how we communicate, learn, consume content, form tastes, create relationships, and more.

There are a variety of risks and benefits associated with AI. The risks include the reduction of human autonomy, attacks on privacy, biased recommendations and more. The benefits include better and faster decision-making, and the saving of time and money.

In order to mitigate the harms and take advantage of the benefits of AI, it is important that individuals understand what it is, how it works, and what it does. Knowledge and understanding can also facilitate better choices and soothe anxieties about unknown systems.

Stage 2: Awareness of Operation

Examples of commonly used systems which are certain or very likely to utilize AI include:

- Voice assistants
- Search engines (of any kind, including those embedded in non-search products)
- Any service which recommends or personalizes content

- Wearable technologies
- Streaming platforms
- Digital or online shopping platforms/apps
- Any service which has a news feed (including news aggregators)
- Social media platforms
- Any system which automatically profiles individuals and places them into groups
- Face-ID and other systems which identify objects

Knowing when an AI is operating is the first step to learning about it.

Stage 3: Prioritization

[Use stages 3, 4 and 5 if you are in a non-time sensitive situation. If you are in a time-sensitive situation, skip to stage 6].

In order to ascertain the urgency and importance of learning about an AI system, individuals may ask themselves the following questions:

- How many times and how many minutes/hours per day do you use the service?
- How/How much does the service affect your decisions, motivations, emotions or beliefs in your everyday life?
- How trustworthy is the organization which provides the service?
- What is the importance of the activity which the service provides to your life and to the health of society?

Individuals should learn about more impactful systems first.

Stage 4: Initial Research, Gaining Knowledge

After selecting a system individuals can start researching it. In doing so it is imperative that they follow good research practices:

- Consider the track-record and reputation of the journalist and/or publisher of each source
- Cross reference all information against at least 2 sources
- Consider the context in which each piece was written
- Consider all information in the light of reason, and experiences, memories and thoughts of using AI systems in the past
- Avoid fringe opinions from insular online communities
- Compile information into a written document to ensure its reusability and structure

In relation to each algorithmic product/service, individuals should focus on three topics at first: basic explanations, features & settings and rights.

Basic Explanations

If the AI system selected is produced by a well-known private organization, individuals should research how said organizations make money. There are many explanations of how firms like Google, Amazon, Facebook etc. profit in journalistic sources.

Afterwards, or if the AI system selected is not produced by a well-known private organization, individuals should research the 'logic' of the algorithm. This includes:

- Its existence
- Its aim
- How its aim is fulfilled

For less well-known systems, individuals may wish to search for generic types of algorithms. For instance, it would be better to search "clothing site recommendation algorithm" than "TopShop algorithm".

Moreover, for practical reasons, it may be better to search “how X algorithm works” than “what is the logic of X algorithm”.

Useful sites for researching the logic of algorithms include:

- Wired
- Techslang
- TowardsDataScience
- EthicalAI.AI (blog)
- SproutSocial

These sites may contain some inaccuracies. As such, individuals should remain vigilant about cross-referencing.

Features & Settings

Many organizations allow individuals to modify their AI’s function. For instance, Twitter allows users to change their feeds from ‘recommended tweets’ to ‘latest tweets’.

As such, it is vital that individuals know the full extent of AI-based services’ features, settings and options.

This can be achieved by searching “how to use X”, watching tutorials, reading users manuals, or simply navigating to the settings page and exploring different choices.

Rights

It is important for individuals to understand their digital rights. This can be achieved by visiting your government’s website, or the website of your nation’s digital/data regulator.

As well as directing individuals to particular areas of concern such as privacy, sustainability, data usage etc., knowledge of rights facilitates their exertion. If individuals live in a country with rights to access or rights to explanation, they should look into exerting them. This is a useful first-hand mechanism for seeing into how AI works.

Stage 5: Delving Deeper

[This stage is optional].

Individuals who wish to learn more about AI systems may be interested in the following topics:

- Human involvement in creating particular systems / the history of AI

Various accounts of the early days of large technology companies are available online or in print.

- The collection and use of personal data

A wide variety of lay sources are available on this subject.

- Types of AI systems

From recommender systems to page-rank algorithms, individuals may wish to gain an understanding of all the processes and outcomes of AI systems.

- Discourse

Individuals may wish to delve into the discourse on technology policy, regulation of AI, or AI ethics. Critical perspectives can be useful for understanding AI and engaging with other people's ideas is an effective mechanism for testing one's own.

[It is worth noting that iterating between prioritization, initial knowledge, and deeper understanding, for different services, will produce 'generalizable' understanding of AI. This will mean that each iteration should be easier than the last].

Stage 6: A Thought Experiment

[Use this stage if you have limited time to consider an AI system before using it].

If individuals need to use an AI system without researching it first, they should consider the following:

- What goals could the organization which created this system have, beyond just providing the service?
- Do you object to the goals of this organization?
- Could the goals of this organization lead to harms occurring?

- If you have objections, are they strong enough to justify not using this system?
- If not, is there anything that you can reasonably change about your use of this system to minimize any harmful effects?

It is worth mentioning that this experiment should never replace research in the long run, and that guesswork about harms and goals will never be exhaustive. Nonetheless, this thought-experiment can be a helpful tool.

Disclaimer

Individuals should note that knowledge of AI cannot cover all systems in existence. Moreover, AI is a fast-changing field, so some information may become obsolete. As such, complacency is ill-advised. Individuals are also asked to use knowledge responsibly. For instance: only manipulate algorithms or 'game' systems, if no one is harmed by doing so.

4.3 Autonomy

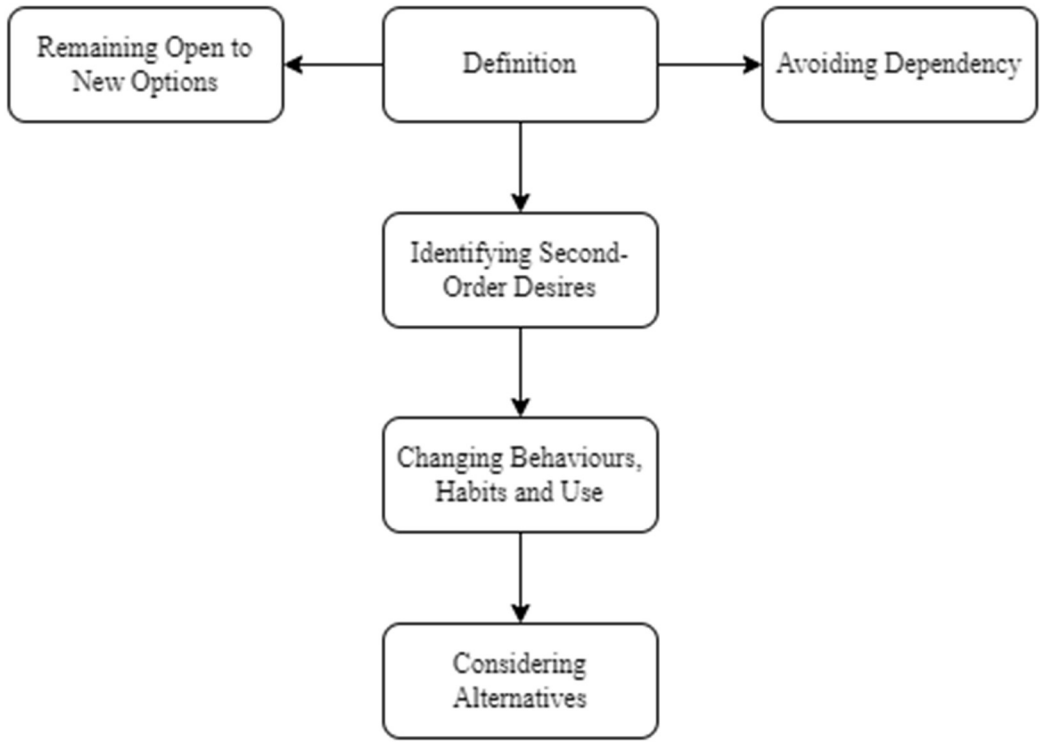

Figure 2: A Framework for the Protection and Enhancement of Autonomy

Figure 2 displays a flowchart representation of the framework for the protection and enhancement of autonomy in relation to AI systems. Each stage of this flowchart is explained in the following section. Importantly, it is assumed that individuals using this framework have already at least partially employed the previous framework on knowledge and understanding.

4.3.1 Definition

Individuals' preconceptions about autonomy may be misguided. As such, they are provided with Dworkin's definition and an illustrative example of it. Interviewee John Zerilli suggested that Dworkin's use of "higher order desires" may be tricky for individuals to grasp. His suggested alternative distinction, between 'wanting' and 'wanting to want', is explained in the framework as well. This provides a more intuitive plainspoken idea for individuals to fall back upon. To incentivise their use of the framework individuals are also informed of the benefits of being autonomous, as detailed in Chapter 2.

4.3.2 Identifying Second-Order Desires

Autonomous individuals must critically reflect on their actions, behaviours and habits (Sumner, 1996). However, individuals will not be naturally inclined to do so, especially in regard to services which they are accustomed to using every day. As such, the framework suggests that individuals should actively identify the second-order desires which they wish to see fulfilled by each service, or type of service, that they currently use. Identification of specific reasons for using specific products, or as Elena Falco, a scholar of AI explainability, put it "starting with what you need something for", was a common trait among interviewed AI ethicists. Interviewees mentioned various changes, from only using Twitter to remain in sync with academic discourse to only using LinkedIn exclusively to search for jobs and research other individuals competing for roles.

Identifying second-order desires need not be a strenuous task. Reflections regarding the usefulness or enjoyment individuals want to derive from a service will be intuitive. Moreover, decisions needn't be made a priori and can be informed by past use of the product (Dworkin, 1988). Furthermore, higher-order desires needn't be based on serious life goals: amusement would be a legitimate second-order desire (Dworkin, 1988). Second-order desires can also evolve over time, as values, behaviours, experiences or the systems themselves change. For instance, as interviewee Lena Bjørlo, a scholar researching consumer autonomy and AI, pointed out: individuals who engage with a field, topic or type of content may find that their need for recommendations lessens as their expertise grows (Sankaran & Markopolous, 2021). It is further worth noting that individuals should not feel restrained by their lists of second-order desires. Exploration, spontaneity and curiosity regarding new options and features should all be encouraged (Dworkin, 1988).

There are many benefits to identifying second-order desires in relation to specific AI services. Most obviously, forcing individuals out of the passive and inert roles which AI-producing entities rely upon facilitates the endorsement of positive habits and the rejection of useless or damaging behaviours. Moreover, it is hoped that individuals who have critically reflected on these matters once will be more reticent to click "accept", "follow" or "keep default settings" in the future. Identifying second-order desires also provides an effective barrier against algorithmic manipulation in the long-term. Indeed, individuals using AI systems for pre-defined reasons will spot deviations from those

reasons more easily. For instance, an individual who explicitly uses Facebook to stay in touch with their family may notice excessive time spent on Facebook Reels more easily than the unthinking normal user. The identification of second-order desires may also mitigate digital exceptionalism. Individuals' behaviours online are often very different to those offline (Calvo et al., 2020). Many of these altered behaviours are facilitated by AI's manipulations and can harm personal autonomy: for instance, 'doomscrolling' or falling into spirals of mindless content consumption. Re-grounding people in non-digital morality and decision-making processes may prevent the most significant deviancies without acquiescing to outdated Luddism. Finally, having individuals use AI systems for specific purposes enshrines the idea of "complementarity", or conceptualisation of AI as a tool or an aid to human intelligence and judgement, rather than an influence or replacement for it (Bjørlo et al., 2021). As well as promoting the use of individuals' judgement this will likely make them more comfortable with the idea of AI.

It is worth noting that this framework avoids informing individuals of which second-order desires should be fulfilled by particular services. This is because different individuals use systems differently. For instance, one interviewed AI ethicist mentioned that they found news feeds and recommender systems useful to "parse content" which was vital for their efficacious consumption of information, while another discussed their commitment to non-AI-based exploration of media which was vital to their sense of authenticity.

4.3.3 Changing Behaviours, Habits and Use

After identifying how AI systems should fulfil their second-order desires, individuals must alter their behaviour accordingly. These changes can occur online or offline and can relate to behaviours, habits or use. Examples are offered in the framework. This practice forms a separate stage in the framework to encourage focused thinking about second-order desires in the previous stage.

4.3.4 Considering Alternatives

Earlier stages of this framework direct individuals in intuitive, focused and likely quite brief processes which can preserve their autonomy. However, after these have been completed individuals may wish to consider AI and autonomy at a finer level of granularity. This includes the consideration of various alternatives rather than individual products and specific effects on autonomy which they may have to research rather than

simple, innate, self-generated desires. Employing nuanced analysis of AI-specific concerns to select or switch products was extremely common amongst interviewed AI ethicists.

To do so individuals should consider the range of alternative products available in each field of service they use (search engines, messaging services, social media etc.). They should judge them against a number of important autonomy-preserving and autonomy-harming criteria, which follow those explained in Chapter 2 of this thesis. The advantages of minimizing autonomy-harms and maximizing autonomy-benefits are self-explanatory. The framework also asks individuals to consider other harms or benefits associated with AI systems, relating to values such as privacy, security, fairness and sustainability (Fjeld et al., 2020). The framework cannot in good faith advocate switching products if autonomy is enhanced, but other values are harmed.

No hierarchy of values can be imposed by the framework. Imposing structure on individuals' desires would counteract the authenticity of their decisions thus undermining many conceptions of the value of autonomy itself (Sumner, 1996). Moreover, it is evident that individuals' priorities often differ. This was made stark in interviews where AI ethicists regularly prioritized different values, and as such, generally avoided different products to one another (barring mass abandonment of Facebook).

4.3.5 Avoiding Dependency

It is important for individuals to avoid becoming dependent upon AI in domains which are important to their lives. Most importantly, they must not allow key skills to atrophy (Varshney, 2020). This is most pertinent when a) a skill is complex, b) the stakes of employing said skill are high, and c) the AI system may not always perform this skill. The framework therefore advises individuals to set calendar reminders to refresh certain skills at regular intervals.

Take the following examples. Forgetting how to aggregate celebrity gossip would not be of concern. Performing this task without social media would be easy due to the prevalence of relevant magazines and television shows. The stakes of not performing this task are likely low. Moreover, TMZ is unlikely to cease reporting entertainment news. However, becoming dependent on an AI to operate a boiler or water-heater would be of concern. Performing this task manually requires knowledge and practice and

lacking heating in one's house could be a significant danger. Moreover, smart home AIs are not installed in every home an individual might live in and may malfunction or fail. As such, this skill must not be allowed to atrophy.

Individuals should also ensure that they retain "input knowledge". These are settings or commands which are given to AIs to automate which may then be forgotten. Returning to the example of the water-heater, an individual might set their smart home's temperature at 20 degrees Celsius. If the AI system malfunctions, it is worth individuals knowing the ideal temperature of their house. The same principle applies to recurring transactions made by banking AIs, repeat orders made by shopping AIs, etc..

4.3.6 Remaining Open to New Options

As detailed earlier in this thesis, AI systems can benefit individuals' autonomy. As such, users must be encouraged to keep an open mind about autonomy-enhancing AI. This point would not have functioned within the stage on considering alternatives since new options needn't arise from field-by-field analysis and will often be discovered by chance. As such, this is the closing remark of the framework.

4.3.7 The Framework for the Protection and Enhancement of Autonomy in Relation to AI

The box below presents the full framework for the protection and enhancement of autonomy in relation to AI, as it would be distributed.

A Framework for the Protection and Enhancement of Individual Autonomy in Relation to AI

Stage 1: Definition

Philosopher Gerald Dworkin defines autonomy as:

The ability "to reflect critically upon … first-order preferences … and the capacity to accept or attempt to change these in light of higher order preferences and values".

In this definition, a first-order preference is a fleeting desire, a split-second instinctive decision, or an impulse, about something which is currently happening. A second-

order or higher-order desire is a more stable, reflective, critical desire, which may be about a first-order desire.

Take the following example. An individual might have the short-term impulse, or first-order preference, to click on a brightly coloured YouTube thumbnail with a catchy title like "You won't believe what happened next". However, they might have a long-term reflective preference, or second-order desire, not to be so easily distracted by these sorts of videos. An autonomous individual recognizes this, reflects, and changes their behaviour accordingly.

To put it another way, an autonomous individual ensures that what they "want" in the short-term, aligns with what they "want to want" in the long-term. As such, they act in accordance with their will and choose the life they want to live. Individuals may feel free to use this distinction instead of Dworkin's if they wish.

Being autonomous is recognized by psychologists to give individuals a feeling of purpose, self-motivation and satisfaction. It is also highly valued by philosophers as an element of wellbeing, a facilitator of personal growth and prerequisite of moral decision-making.

Stage 2: Identifying Second-Order Desires

Individuals should critically reflect on the second-order desires that are being, or should be, satisfied by each particular AI-based service they use. In other words, they should identify what they 'want to want' from a particular service in the long-run. For instance, one might want to use Facebook to fulfil the following second-order desires:

- The desire to stay in touch with family
- The desire to stay up to date with events in a local area
- The desire to be reminded of friends' birthdays
- The desire to do all of the above quickly and easily, without the need for constant checking in

Second-order desires do not need to be based on serious life goals. 'Amusement' is a legitimate second-order desire. They can also be based on experience of using a service, rather than being generated before initial use. Second-order desires may also change over time.

It is worth repeating this process whenever you adopt a new service.

Stage 3: Changing Behaviours, Habits and Use

After identifying second-order desires associated with AI systems, individuals must consider what they could change about their behaviour to better fulfil those desires.

Changes can occur online, and relate to use of services, altering settings, etc. Equally they can occur offline and involve avoidance of certain services, mindfulness of time spent using certain products, etc.

With this said, individuals should not feel constrained to always act in accordance with their pre-designated second-order desires. Exploration, spontaneity, change and curiosity regarding new options and features must not be sacrificed.

Stage 4: Considering Alternatives

[Complete this stage if you wish to analyse autonomy and AI at a finer level of granularity. If not, skip to stage 5].

Next, individuals should consider the wider landscape of AI systems. Here, individuals may wish to consider not only how algorithmic products or services can fulfil their second-order desires, but how they affect their *ability* to fulfil second-order desires as well. In doing so, they should ask the following questions of each AI in a given field (e.g., search engines, social media etc.):

1. Does each AI system promote individual autonomy?
   - How well does/would this product fulfil your second-order desires?

- Does this product save time, energy or money?
- Does this product facilitate self-nudging? (Some AI systems can remind individuals to complete certain tasks. This would aid one's *ability* to be autonomous.)
- Does this product provide access to novel options, which other systems do not?

2. Does each AI system harm individual autonomy?

- Is the algorithm manipulative?
- Do you find yourself using it in ways that you did not originally intend?
- Does it have an unintended effect on your decision-making, emotional state, or thoughts?

- Do elements of the system's design indirectly affect your preferences? (For instance, if you used an online shopping website, which designated low-cost shoes to be between £20 and £50, this may lead you to believe that spending up to £50 on shoes is frugal).

- Do you feel your choices being limited or bounded by an algorithm's recommendations?

With this said, products should not only be considered on the basis of autonomy. As such, individuals should give a little time to researching other factors.

3. Are there other harms or benefits associated with the system?

- Privacy, security, fairness and sustainability may all be issues of concern. For guidance on researching these sorts of ideas, please return to the framework on knowledge and understanding.

Importantly, this framework does not impose a hierarchy of values. It is up to individuals to weigh pros and cons.

Stage 5: Avoiding Dependency

It is important for individuals not to become dependent upon algorithms in their everyday lives. As such, they must retain important skills. Characteristics of a skill or ability which deem it important are as follows:

- The skill is complex
- The stakes of employing the skill are high
- The AI system is not always going to perform the skill

It is also important for individuals to retain "input knowledge". This is information which individuals give to algorithmic systems, but then often forget about due to the automatic completion of the relevant tasks. For instance, individuals should remember the existence and details of repeat shopping orders, recurring bank transactions, etc.

Setting calendar reminders to refresh skills and knowledge at regular intervals may be useful for avoiding dependency.

Stage 6: Remaining Open to New Options

AI can enhance individual autonomy as well as harming it. AIs can save time, energy and money; they can provide new alternatives; they can provide effective personalised recommendations and useful information. As such, it is important for individuals to remain open to new AI systems which could enhance their autonomy.

5. Limitations

There are a number of limitations of this research. For instance, a number of constraints on the potential efficacy of this thesis stem from its concentrated focus. Its purely individualistic paradigm may lead to the neglect of other stakeholders (Ananny & Crawford, 2018). This could be especially damaging for stakeholders without voices, such as the environment or children, as interviewee Fabrice Muhlenbach pointed out. Similarly, this thesis' consideration of two key values may lead to the neglect of others. Individuals must consider privacy, fairness, security, bias and sustainability, and trade-offs between them, when making decisions. Moreover, this framework's focus on grounded issues in individuals' everyday lives may lead users to neglect wider societal issues. Technology companies' political and economic power; political manipulation online; concerns around surveillance; criminality and scams should all be considered in relation to AI. Although this thesis' narrow focus was necessary for practical reasons, and although the frameworks do raise the importance of other stakeholders, values, trade-offs/decisions and discourse at the most pertinent junctures, these issues must be accounted for.

Additionally, several constraints relate to individuals' potential use of the frameworks. Most importantly, individuals may identify second-order desires which are not in their long-term best interests. Returning to a previous example, an individual who wants to use Facebook to stay in touch with their family, may not actually benefit from this behaviour if their relationships with family members are, or become, harmful. This is an intractable philosophical problem, but a problem, nonetheless. Individuals may also outrightly disregard key elements of the frameworks; fail to keep recommended habits; become scared or angered by information they find; or feel that the frameworks are condescending. Furthermore, some individuals may struggle to employ these frameworks due to the contexts of their lives: health, intelligence, socio-economic background and age may all hold people back. Effort was made to create clear, concise, comprehensible, accessible resources with positive impacts. However, for some, this thesis may have fallen short.

A further limitation of this thesis is its reach. If ethical frameworks are created to help individuals, and no individuals see it, then what good has been done? As such, the findings of this thesis must be distributed after its submission. The presented ethical frameworks can easily be adjusted and reformatted into web-resources or social media

posts; moreover, it is hoped that the findings of this thesis may be amenable to publication either in an academic journal or in the press.

From an alternative perspective the adoption of these frameworks could potentially pose risks. For instance, much of the framework for 'knowledge and understanding' is based on a modest degree of research ability (being able to use search engines, read journalistic articles etc.). Although the accessibility and practicality of this thesis would have a positive effect on digital divides, its dependence on research may widen them in the long run. Another risk is posed by the possibility of AI ethics 'from below' being used by corporations to shift blame from themselves to users, in a similar pattern to other causes (for instance, environmentalism). These are highly speculative concerns but should be acknowledged by any future researchers.

Despite these limitations this thesis' impact should be positive. It provides an effective and comprehensible set of ideas and practices for individuals to utilise, to protect and enhance their autonomy, and increase their knowledge and understanding, when interacting with AI. Moreover, its frameworks are rooted in respected philosophy, a firm understanding of the AI ethics literature, and interviews with AI ethicists, ensuring its sound theoretical grounding and practical feasibility. If applied, these frameworks will help individuals to regain their agency, in a field where so many are consistently confused about or unaware of the harms they face and benefits they overlook.

6. Conclusion

AI mediates individuals' social environments, tastes, identities and livelihoods. As such, it is critical that they are informed as to the potential harms and benefits of AI systems. Despite this, the current AI ethics literature places focus only on those who have agency over how AI functions, leaving those who have agency over how AI systems are experienced, defenceless. This study has challenged this status quo, displaying how AI ethics can be conceptualized and operationalized for the individual, and providing a practical framework for people to employ, based on interviews with AI ethicists, rigorous philosophical theories and the AI ethics literature.

This thesis' contribution of AI ethics 'from below' does not mean to replace top-down AI ethics. Indeed, 'top-down' and 'bottom-up' approaches can complement and reinforce one another; greater popular concern may provide a mandate for regulation, for instance. With this said, it is strongly hoped that this thesis serves as a starting point, or perhaps proof-of-concept, for a novel and feasibly potent avenue of AI ethics research and practice. If expanded, AI ethics 'from below' can provide a social workaround for mitigating AI's risks and promoting its benefits in the real world: one which cuts past wilfully lax corporate policies; past ethics as a marketing exercise; past slow regulatory practices. AI ethics 'from below' can engage psychologists, philosophers, sociologists, engineers and designers alike, in circumventing traditional hurdles to the field, and instigating real change in people's lives. It is this thesis' hope that such a field might bloom.

Bibliography

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: adapting the TSE framework to big data. Journal of Survey Statistics and Methodology, 8(1)

Ananny, M., Crawford, K. (2018). Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, Vol 20 (3)

Andow, J. (2016). Qualitative tools and experimental philosophy. Philosophical Psychology, 29 (8).

Applin, S. A., & Fischer, M. D. (2015). New technologies and mixed-use convergence: How humans and algorithms are adapting to each other. In 2015 IEEE International Symposium on Technology and Society.

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. International Journal of Information Management, 62

Audi, R. (2011). Epistemology: A contemporary introduction to the theory of knowledge. New York: Routledge.

Ausloos, J., Dewitte, P., Geerts, D., Valcke, P., Zaman, B. (2018). Algorithmic transparency and accountability in practice. Computer-Human Interactions.

Balkin, J.M. (2015). Information fiduciaries and the first amendment. UCDL Rev., 49

Baudouin, V., Bloch, I., Bounie, D., Clemencon, S., d'Alche-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., Parekh, J., (2020). Flexible and context-specific AI explainability: a multidisciplinary approach. Computers and Society.

Baum, S.D. (2020). Social choice in artificial intelligence. AI & Society.

Baumberger, C. Beisbart, C. & Brun, G. (2017). What is understanding? An overview of recent debates in epistemology and philosophy of science. Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science.

Bjørlo, L.V., Moen, O., Pasquine, M. (2021). The role of consumer autonomy in developing sustainable AI: a conceptual framework. Sustainability, 13.4

Blank, G. (2017) Online research methods and social theory. In The Sage Handbook of Online Research Methods. Sage Publications.

BonJour, L. (2017). The dialectic of foundationalism and coherentism. The Blackwell Guide to Epistemology. Wiley.

Brinkman, S., Kvale, S. (1996). Thematizing and designing an interview study. In InterViews: learning the craft of qualitative research interviews. Sage Publications.

Brinkman, S., Kvale, S. (2019). Epistemological issues of interviewing. In Doing interviews. Sage Publications.

Burke, P. (2015). The French historical revolution: the Annales school, 1929-2014. Cambridge: Polity Press.

Calvo, R.A., Peters, D., Vold, K., Ryan, R. (2020). Supporting human autonomy in AI systems: a framework for ethical enquiry. In Burr, C., Floridi, L., Ethics of Digital Wellbeing, Springer.

Candidate 1061912, (2022). Digital interviewing summative paper. Oxford Internet Institute.

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and legal challenges. Philosophical Transactions of the Royal Society.

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., & Taylor, L. (2018). Portrayals and perceptions of AI and why they matter, Royal Society & the Leverhulme Centre for the Future of Intelligence.

Chiodo, S. (2022). Human autonomy, technological automation (and reverse). AI & Society, 37(1)

Christman, J. (2020). Autonomy in Moral and Political Philosophy. Stanford Encyclopaedia of Philosophy.

Cherry, D., LaRock, T. (2014). The basics of digital privacy: simple tools to protect your personal information and your identity online. Massachusetts: Syngress.

Diakopolous, N. (2020). Transparency. In the Oxford handbook of ethics of AI. Oxford: Oxford University Press.

Dignum, V. (2017). Responsible autonomy. arXiv:1706.02513

Dworkin, G. (1988). The theory and practice of autonomy. Cambridge: Cambridge University Press.

Edwards, R., Holland, J. (2013) How have qualitative interviews developed?. In What is Interviewing. Bloomsbury.

Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. AI & ethics, 1

European Parliamentary Research Service. (2020). The ethics of artificial intelligence: issues and initiatives.

Fast, E., Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. In Proceedings of the AAAI conference on artificial intelligence, Vol 31

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication.

Flick, U. (2009). Sampling. In An introduction to qualitative research. Sage Publications.

Floridi, L. (2002). What is the philosophy of information?. Metaphilosophy, 33(1-2)

Floridi, L. (2010). Information: a very short introduction. Oxford: Oxford University Press.

Floridi, L. (2014). The 4th revolution. Oxford: Oxford University Press.

Floridi, L. (2017). The digital's cleaving power and its consequences. Philosophy and Technology, 30

Floridi, L. (2018). Semantic capital: its nature, value, and curation. Philosophy & Technology, 31 (4)

Floridi, L., Cowls, J. (2019). A unified framework for five principles for AI in society. Harvard Data Science Review, 1.1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4)

Floridi, L., Cowls, J., King, T.C., Taddeo, M. (2020). How to design AI for social good: seven essential factors. Science and Engineering Ethics, 26

Floridi, L. (2021). Translating principles into practices of digital ethics: Five risks of being unethical. In Ethics, Governance, and Policies in Artificial Intelligence. Springer.

Fumerton, R. (1990). Metaepistemology and scepticism. In Doubting. Dordrecht: Springer.

Gabriel, I. (2020). Artificial intelligence, values and alignment. Minds and Machines.

Grimm, S. (2006). Is understanding a species of knowledge? The British Journal for the Philosophy of Science, 57(3)

Grimm, S. (2011). Understanding. In The Routledge Companion to Epistemology. New York: Routledge.

Hagendorff, T. (2020). The ethics of AI ethics: an evaluation of guidelines. Minds and Machines, 30(1)

Hagerty, A., Rubinov. I. (2019). Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. Computers and Society.

Hand, M. (2014). From cyberspace to dataverse: trajectories in digital social research. In Big data: qualitative approaches to digital research. Emerald Group Publishing.

Harvey, W.S. (2011). Strategies for conducting elite interviews, in Qualitative Research, 11 (4)

Hatch, J. A. (2002). Doing qualitative research in education settings. State University of New York Press.

Hayek, F.A. (1960). The constitution of liberty. University of Chicago Press.

Heald, D. (2006). Varieties of transparency. In: Hood C and Heald D (eds) Transparency: The Key to Better Governance? New York: Oxford University Press.

Hermann, I. (2020). Beware of fictional AI narratives. Nature Machine Intelligence, 2(11)

Hewson, C. (2014). Qualitative approaches in internet mediated research: opportunities, issues, possibilities. In The Oxford handbook of qualitative research. Oxford: Oxford University Press.

Hill, T.E. (2013). Kantian autonomy and contemporary ideas of autonomy. In Sensen, O. (ed.) Kant on Moral Autonomy. Cambridge University Press.

Hove, S. E., & Anda, B. (2005). Experiences from conducting semi-structured interviews in empirical software engineering research. In 11th IEEE International Software Metrics Symposium.

Howlett, M. (2021). Looking at the field through a zoom lens: methodological reflections on conducting online research during a global pandemic. Qualitative Research, 21(1)

Hu, Q., Lu, Y., Pan, Z., Gong, Y., Yang, Z. (2021). Can AI artifacts influence human cognition? The effects of artificial autonomy in intelligent personal assistants. International Journal of Information Management, 56

Hyman, J. (1999). How knowledge works. The Philosophical Quarterly, 49(197)

Ireni-Saban, L., & Sherman, M. (2020). Incorporating intersectionality into AI ethics. In Democracy and Fake News. Routledge.

James, N., Busher, H. (2011). Epistemological dimensions in qualitative research: the construction of knowledge online. Online Interviewing. Sage Publications.

Jia, H., Wu, M., Jung, E., Shapiro, A., & Sundar, S. S. (2012). Balancing human agency and object agency: an end-user interview study of the internet of things. Proceedings of the 2012 ACM Conference on Ubiquitous Computing.

Jobin. A., Ienca, M., Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1

Johnson, D.G., Verdicchio, M. (2017). Reframing AI discourse. Minds & Machines, 27

Kant, I., Wood, A., Schneewind, J.B. (2002). Groundwork for the metaphysics of morals. Yale University Press.

Karray, F., Alemzadeh, M., Abou Saleh, J., Arab, M.N. (2017). Human-computer interaction: overview on the state of the art. International Journal on Smart Sensing and Intelligent Systems, 1(1)

Kazmer, M.M., Xie, B. (2008). Qualitative interviewing in internet studies: playing with the media, playing with the method. Information, Community and Society.

Kemper, J., Kolkman D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. Information, Communication & Society. 22:14

Kop, M. (2021). EU artificial intelligence act: the European approach to AI. Transatlantic antitrust and IPR developments.

Kvanvig, J. (2003). The value of knowledge is external to it. In The value of knowledge and the pursuit of understanding. Cambridge: Cambridge University Press.

Kvanvig, J.L. (2017). Understanding. In The Oxford handbook of the epistemology of theology. Oxford: Oxford University Press.

Langer, M., Baum, K., Hartmann, K., Hessel, S., Speith, T., & Wahl, J. (2021). Explainability auditing for intelligent systems: a rationale for multi-disciplinary perspectives. IEEE 29th International Requirements Engineering Conference Workshops (REW).

Langer, M., Hunsicker, T., Feldkamp, T., Konig, C.J. & Grgic-Hlaca, N. (2022). 'Look! It's a computer program! It's an algorithm! It's AI!': Does terminology affect human perceptions and evaluations of algorithmic decision-making systems?. Computer-Human Interactions. Conference on Human Factors in Computing Systems. New York: ACM.

Langer, M., Oster, D., Speith, T. & Hermanns, H., (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence.

Lhérisson, P. R., Muhlenbach, F., & Maret, P. (2017). Fair recommendations through diversity promotion. International Conference on Advanced Data Mining and Applications. Springer.

Lieberson, S. (1987). Making it count. University of California Press.

Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human-AI interaction. Journal of Computer-Mediated Communication, 26(6)

Mao, C., Koide, R., Brem, A., Akenji, L. (2020). Technology foresight for social good: social implications of technological innovation by 2050 from a global expert survey. Technological Foresight and Social Change.

Martınez-Miranda, J., & Aldea, A. (2005). Emotions in human and artificial intelligence. Computers in Human Behaviour, 21 (2)

McMillan, D., Brown, B. (2019). Against ethical AI. Proceedings halfway to the future symposium.

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. AI & Society, 35(4)

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

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1.11

Mökander, J., Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. Minds and Machines, 31(2)

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of AI ethics. Minds and Machines, 31(2)

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., Floridi, L. (2021). Operationalising AI ethics: barriers, enablers and next steps. AI & Society

Morley, J., Morton, C., Karpathakis, K., Taddeo, M., & Floridi, L. (2021). Towards a framework for evaluating the safety, acceptability and efficacy of AI systems for health: an initial synthesis. arXiv.

Muhlenbach, F. (2020). A methodology for ethics-by-design AI systems: dealing with human value conflicts. IEEE International Conference on Systems, Man, and Cybernetics (SMC).

Nel, A. L., & Carroll, J. (2017). Ethics assessment via game play? In 2017 IEEE Global Engineering Education Conference.

Olsson, E.J. (2011). The value of knowledge. Philosophy compass, 6

Patton, M. Q. (2002). Qualitative research & evaluation methods. Sage Publications.

Plakias, A. (2015). Experimental philosophy. In The Oxford Handbook of Topics in Philosophy. Oxford: Oxford University Press.

Preece, A. (2018). Asking 'why' in AI: explainability of intelligent systems – perspectives and challenges. Intelligent Systems in Accounting, Finance and Management, 25(2)

Pritchard, D. (2007). Recent work on epistemic value. American Philosophical Quarterly, 44(2)

Pritchard, D. (2009). Knowledge. Basingstoke: Palgrave Macmillan.

Pritchard, D., Turri, J., Carter, J.A. (2018). The value of knowledge. The Stanford Encyclopaedia of Philosophy.

Prunkl, C. (2022). Human autonomy in the age of artificial intelligence. Nature Machine Intelligence, (4)

Recchia, G. (2020). Investigating AI narratives with computational methods, in AI Narratives: A History of Imaginative Thinking about Intelligent Machines, 382

Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data & Society, 7(2)

Richter, F. (2021). Ethics of AI as practical ethics. IEEE International Symposium on Technology and Society (ISTAS).

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., Florid, L. (2021). The Chinese approach to artificial intelligence: an analysis of policy, ethics and regulation. AI & Society, 36

Roff, H. (2019). Artificial intelligence: power to the people. Ethics & International Affairs, 33(2)

Saldana, J. (2015). The coding manual for qualitative researchers. Sage Publications.

Sankaran, S., Markopolous, P. (2021). 'It's like a puppet master': user perceptions of personal autonomy when interacting with intelligent technologies. Proceedings of the 29th Conference on User Modelling, Adaptation and Personalisation.

Sankaran, S., Zhang, C., Gutierrez Lopez, M., & Väänänen, K. (2020). Respecting human autonomy through human-centered AI. Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society.

Sartre, J.P. (1956). Being and nothingness: an essay on phenomenological ontology. New York: Philosophical Library.

Schiff, D., Biddle, J., Borenstein, J., Laas, K. (2020). What's next for AI ethics, policy and governance? A global overview. Proceedings of the AAAI/ACM conference on AI, Ethics and Society.

Schlosser, M.E. (2019). Dual-system theory and the role of consciousness in intentional action, in Feltz, B., Missal, M., Sims, A. (eds.), Free Will, Causality and Neuroscience, Brill Editions.

Seger, E., Avin, S., Pearson, G., Briers, M., Heigeartaigh, S. Ó., Bacon, H. (2020). Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world. Alan Turing Institute.

Shin, D. (2021). The effects of explainability and causability on perception, trust and acceptance: implications for explainable AI. International journal of human computer studies, 146

Skinner, Q. (1969). Meaning and understanding in the history of ideas. History and Theory, 8(1)

Steup, M., Ram, N. (2020). Epistemology. The Stanford Encyclopaedia of Philosophy.

Stratton-Lake, P. (2013). Rational intuitionism. In Crisp, R. The Oxford handbook of the history of ethics. Oxford: Oxford University Press.

Sumner, L. (1996). Welfare, happiness and ethics. Oxford: Clarendon Press.

Taddeo, M., Floridi, L. (2018). How AI can be a force for social good. Science, 361

Theodorou, A. (2019). AI governance through a transparency lens. University of Bath.

Theodorou, A., Dignum, V. (2020). Towards ethical and socio-legal governance in AI. Nature Machine Intelligence, 2 (1)

Tracy, S.J. Qualitative quality: eight 'big tent' criteria for excellent qualitative research, Qualitative Inquiry, 16(10)

Turri, J., Alfano, M., Greco, J. (2021) Virtue epistemology. The Stanford Encyclopaedia of Philosophy.

Van Haitsma, M. (2009). Key informant interviews.

Varshney, L.R. (2020). Respect for human autonomy in recommender systems. arXiv:2009.02603

Wachter, S. (2018). The GDPR and the Internet of Things: a three-step transparency model. Law, Innovation and Technology, 10(2)

Wacks, R. (2015). Privacy: A very short introduction. Oxford: Oxford University Press.

Weiser, M. (1999). The computer for the 21st century. ACM mobile computing and communications review, 3(3)

Wischmeyer, T. (2020). Artificial intelligence and transparency: opening the black box. In regulating artificial intelligence. Springer.

Yang, G. Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kuman, V., McNutt, M., Merrifield, R., Nelson, B.J., Scassellati., B., Taddeo, M., Taylor, R., Veloso, M., Wang, Z.L., Wood, R. (2018). The grand challenges of science robotics. Science Robotics, 3(14)

Zagzebski, L. (1996). Virtues of the mind: an inquiry into the nature of virtue and the ethical foundations of knowledge. Cambridge: Cambridge University Press.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic decision-making and the control problem. Minds and Machines, 29 (4)

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: is there a double standard?. Philosophy & Technology.

Zerilli, J. (2021). A citizen's guide to artificial intelligence. MIT Press.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. Journal of Information Technology