ARTIFICIAL BELIEFS

Pål Rostad Stokholm

Kandidatnr:101

Masterprogram:

Fil4091: Mastergradsessay i filosofi.

Supervised by

Sebastian Watzl, Associate professor

Universitetet i Oslo

Institutt for filosofi, ide- og kunsthistorie og klassiske språk

Oslo,

15.12.2020

Acknowledgements.

I want to thank my supervisor Sebastian Watzl, for his advice and support. The talks we had during the Covid19 pandemic has been memorable and motivating through these strange and difficult times. Without your mentoring, I would never have found my focus. I also want to thank my parents for the unbelievable courage and strength they have shown in this period as well. My father, who inspires my mind, and my mother who warms my heart.

I want to thank my friends at IFIKK and my friend Eric Ian Andersen, whose talks have always inspired and grounded me.

Oslo,

December 2020.

Abstract.

This essay concerns the epistemology of beliefs formed on the basis of the output of AI systems. To do this, I follow the definition given by the European council's High-Level Expert Group on Artificial Intelligence (AI HLEG). I discuss some of the epistemically problematic implications of that definition and propose a tentative epistemological taxonomy of AI. This taxonomy is based on the dynamic between opacity and autonomy, on a scale of the total complexity of systems that acts as an external source of knowledge. By using this taxonomic system, I then suggest possible theories of justification used for other sources of knowledge that may be suited for a variety of AI depending on where it fits in the taxonomy. In the end, I conclude that social elements are the most important part of a plausible theory of justification for the reliance, or trust, in AI.

Contents

1. Introduction.

The emergence of Artificial Intelligence (AI) is changing how we do almost everything in some aspect or another. It is changing many of the fundamental ways we operate in business, culture, politics, academia and science. As its implementations continue to integrate itself into our everyday lives, it is affecting more than just how we conduct our practical lives, but also our mental lives. Intelligence is - for the first time, available on-demand.

This happens so frequently that we are often unaware of how often AI not only acts as the source of our beliefs but also how we justify them. However, it is surprising that there has not been a more widespread inquiry into the potential epistemological issues surrounding our reliance on AI as a soruce of knowledge. What kind of source of knowledge does AI represent? Is it an instrument, or is it the emergence of a new kind of rational agent? The goal of this essay is to discuss these problems and discuss some of the plausible solutions for the epistemic problems that face AI now, and in the near future.

1.1 What is the epistemic status of beliefs based on the output of AI?

In this section, I will introduce some of the ideas about the epistemology of artificial intelligence (AI) that will be discussed in this essay.

When I have a belief based on the output of an AI system, that belief has an epistemic status that can be either positive or negative depending on whether I have a good or bad theory of justification for that belief. These theories often consist of conditions that must be satisfied for the belief to be justified. If the conditions are satisfied, we have a justification for why we believe what we believe. If the belief also happens to be true; then we have a justified true belief, which is often considered necessary for claims of knowledge. This essay, however, will focus on

how the epistemic features of AI affect the way we can justify our belief in the output, or testimony of AI, rather than what is required to make claims of knowledge based on such beliefs.

Or in other words, for a subject S to have a justified proposition p based on the output of AI; what conditions are plausible for us to say that S is justified to believe p? In this section, I will spend some time discussing the problems defining AI and how changes in its epistemic features may affect how we justify our belief in its output or testimony. Following the definition of the European Commission's Artificial Intelligence High-Level Expert Group (AI HLEG), I will argue that the relevant epistemic features for justification of AI beliefs are found in the degree of epistemic opacity and autonomy of the AI system, which is also relative to the systems total complexity. In section one, I will argue that by focusing on these epistemic features, we can have an epistemic taxonomy of AI that does not solely depend on the technical features of the AI system – although the two are often connected. In section two, I will discuss plausible justifications of AI-based beliefs that is relative to how the dynamic between these 3 dimensions affects its epistemic features.

1.1.2 Basic reasons for belief in external sources.

In this section, I will mention some of the most salient reasons for belief in external sources of knowledge. Let us first ask, what are some of the common reasons for believing in a proposition given by an external source? In most cases, the conditions for why we should believe a proposition to be true depends on – and is relative to, how the proposition was delivered to the subject. The means of delivery matters in the sense that we are entitled to a degree of scepticism depending on how the propositional content was delivered to us. For instance, if a proposition came to me from memory, like my home address, then my memory should only be trusted on the condition that I believe that I have a well-functioning memory. Sometimes the deliverance is made by someone else; for instance, if the proposition p was delivered by someone's testimony, then the belief that p is true depends on how trustworthy the testimony is deemed to be. In other cases, like with instruments, the belief delivered by the instrument is justified by the instrument's designed reliability and lack of apparent malfunctions or flaws that could defeat my entitlement to assume its reliable.

1.1.3 AI as different sources of knowledge.

Here, I will introduce the notion of AI as a term for potentially different kinds of sources and how that affects the way we can justify our belief in it.

To analyse the epistemic status of beliefs based on the output of AI, we need to define what kind of source AI is if we are going to establish the conditions that must be satisfied for our reliance on it. Some sources are naturally deemed more reliable than others and thus we have different standards for each respective kind of source. If we are supposed to accept AI as a primary source of knowledge, we must be sure of what it is. This is not as easy as it might seem with AI, both from a technical and metaphysical point of view.

On the most basic level, identifying a source of knowledge implies knowing if the source is an object, a subject, or just something you have inferred yourself. But when it comes to the ambiguous concept of artificial intelligence, the answer is not always as easy as it might seem. It seems that the more complicated AI becomes the more questions arise about exactly what it is. The answer seems to be that AI can take many forms and thus represent fairly distinct sources of knowledge that require different epistemological justifications.

One of the first distinctions that come to mind is whether AI is something that can simply make reliable calculations or indicate propositions – or if it is something that can "tell" us propositions. This is not to say that if an AI can be deemed to "say" things, it is necessarily testifying on par with a human, but that it has a level of epistemic agency that is distinct from what we have previously seen in objects. Something like an artificial animal that can act as a rational agent without having a conscious status that is associated with that kind of ability. Depending on this capacity, the conditions for justification will naturally change.

1.1.3.1 Mistaking what kind of source of knowledge AI is.

If AI can represent different sources of knowledge, how can we determine what source it is? If it can represent external sources, ranging from simple instruments to rational agents, epistemic subjects and epistemic agents, what epistemic features of AI is most salient to you? Think about

it this way; would you be more inclined to believe in the reliability of something that cannot make judgements, or, something that delivers or even "tells" you things? How about something that can give reliable testimony, but at the same time unable to provide sufficient justifications for its deliverances? Intuitively, these differences also affect how to responsibly justify reliance on them, and thus we may encounter some serious problems if mix up these different epistemic abilities or features.

In those difficult situations, perhaps some would prefer the objective output of an instrument, while in other situations, they might prefer to be "told" what to believe by some reliable agent. Neither being told what to believe nor reading facts from some instrument is necessarily considered more preferable depending on the context of use. For instance, most people would rather be diagnosed by a doctor, who relies on both tangible facts and intangible experiences, than the sole output of an unfamiliar AI medical diagnosis system. On the other hand, if we are taking an ice bath in the middle of January, I'd be more inclined to believe the temperature indicated by my AI-based weather app more than my friend's testimony of the water being "pretty mild" – even if I deem his testimony reliable. This may be an obvious point to make, but if we compare this to how AI works – the line between AI's instrument-based output and its self-made consideration about something it has "learned", is not always easy to distinguish. What follows, is that the complex (and often opaque) processes behind the output of AI further complicate the analysis of AI output as it is hard to determine how much is caused by its intellectual capacity or information gathered by its systems instruments like actuators and sensors. For instance, the temperature indication of the weather app may be caused indirectly by its intellectual estimates, or by a direct feed from remote thermostats close to my location.

1.2 Defining AI

In this section, I will discuss the difficulties of defining AI and why I have chosen the HLEG definition. Furthermore, I will argue that the HLEG definition is not without its problems and some of the epistemological implications that follow those problems.

Through the years, a number of attempts have been made to define AI, without any of them becoming philosophical canon. The broad nature of AI has proven to be quite difficult to pinpoint and so a final definition has remained quite elusive from all kinds of definitions. The attempts include dozens, if not hundreds, of definitions that come from a wide array of respectable institutions like The US department of defence (Defense Innovation Board, 2020), Organisation for Economic Co-operation and Development (Berryhill, Kok Heang, Clogher, & McBride, 2019) and The World Economic Forum (Herweijer & Waughray, 2018), as well as AI experts like Russel and Norvig (Russell & Norvig, 2009). However, all of these definitions seem to be lacking universal consensus.

Fortunately, the European Commission recently set up the High-Level Expert Group on Artificial Intelligence (AI HLEG) for the purpose of developing a working definition for the European Union (EU). The group consists of 52 world-leading experts on AI, and their definitions, taxonomy and work on the *"trustworthiness of AI"* seems to be more than adequate for the purpose of this essay. This definition will therefore stand central to this thesis as the grounding definition of AI. But as we will see, their final definition is not without its problems.

1.2.1 The final HLEG definition

On 8 April 2019, The European Commission released their definition that the AI HLEG had been working on since June 2018, after an initial release of their ethics guideline for trustworthy AI the same year. This definition goes as follows:

> *"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their*

> *behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)." (High-Level Expert Group on AI, 2019, p. 9).*

Although this is a complex definition, we can see that HLEG establishes certain conditions that are necessary for something to be AI. The most interesting epistemological feature of this definition is that AI must be:

1. Something that decides the best action(s) to achieve its goal(s).

Of these conditions given by HLEG, this stands out as perhaps the most epistemologically interesting one, as according to HLEG, AI is something that necessarily makes the best decision(s) to achieve its goal(s). If this is true, then remarkably AI must be completely reliable if its goals align with those who use it. AI is thus reliable (or trustworthy as HLEG may say) just by definition, and thus justification of AI-based beliefs can be made by analytic apriori reasoning. That is to say, a part of the meaning of the term "AI" is that finds the *best* action(s) to achieve its goal(s). As a consequence, any malfunctioning AI, or even AI that only makes good decisions – but not the best – is *not* AI.

This seems like a rather problematic (and presumably unintended) consequence of HLEG's definition for a couple of reasons. For one; epistemologically speaking such bold claims seems perhaps a little *too* convenient as it does not leave any room for scepticism about reliance or trust in AI output (or AI testimony). According to HLEG, AI is reliable (or trustworthy) because AI is reliable. Justifications of belief in AI reliability or trustworthiness are thus circular if we follow this definition. This is an epistemological non-starter and is impossible to accept if we want to make serious considerations about beliefs in AI's reliability.

Secondly, since there are currently no infallible AI systems, what follows from HLEG's definition is that AI does not yet exist. So it seems, if we accept this definition, we are stuck with an unwanted circular theory of justification of belief in the reliance of AI, and references to a kind of AI that does not yet exist.

1.2.2 The preliminary HLEG definition

However, HLEG also provides a preliminary definition that avoids unrealistic standards of AI.

> *"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)." (High-Level Expert Group on AI, 2019, p. 1)*

Here, the only two necessary conditions are that AI are *"systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals"* and that it can be based on both software and hardware. Notice that this definition avoids the unwanted condition 3 that caused our philosophical impasse, so to serve our present purposes this will be the definition moving forward. Some may say that this is perhaps is a little *ad hoc*, but I'd say the same can be said to a greater extent about condition 3 in the first definition, so I will leave it at that and move forward.

Another thing to notice with these definitions is that both the final and the preliminary definitions stipulate that AI is something that possesses, at a minimum, *some degree of autonomy*. As this concept of autonomy seems to be an essential characteristic of AI, its implications should be made explicit. But what does "autonomy" mean in the context of AI? Next up, I will spend some time discussing the meaning of autonomy in relation to AI and how it

can often lead to opaque systems. I will also argue that these two elements serve as essential factors when making proper justifications of AI-based beliefs.

1.2.3 AI output as something to trust.

The issues discussed in this essay are often related to the topic of when we are justified to believe in the reliance on AI. However, the AI HLEG often talks of AI in terms of trust. This raises the question – is AI something that can have the capacity to be trusted?

Furthermore, when we look at the different relationships we have in terms of reliance on these different sources, it becomes apparent that the common idea of "trustworthy AI" is something that is becoming relevant in the current discourse as discussed by the AI HLEG in their report: Ethics Guidelines for Trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019). However, epistemology conventionally talks about sources of knowledge in terms of reliability and not "trustworthiness". Some, like Mark Ryan (Ryan, 2020) and Margit Sutrop (Sutrop 2019) have rejected this notion and argued that trust is a social phenomenon grounded in emotions that does not make sense in relation to AI. As they see it, AI simply does not have the capacity to be trusted. Indeed, AI is often anthropomorphized and the notion of trust may be an inappropriate way of understanding any epistemic reliance on an object like artificial intelligence. In philosophy, there is a clear distinction between trust and reliance, yet the notion of trustworthy AI has recently been recognized by the European Commission's High-Level Expert Group on AI (HLEG) as something that *can* - and *should* exist. If the HLEG is right in this assessment, then trust should factor into how we analyse beliefs sourced from artificial intelligence, but how are we supposed to make epistemological sense of "trustworthy" AI? If AI is something to be discussed in terms of trust, perhaps the output of AI is better understood as a source of testimony, which I will discuss later in section 2.4.3. after we have followed the implications of the AI taxonomic categories I suggest in section 2.1.

1.2.4 The difficulties of predicating AI output properly.

In this section, I will introduce some of the problems of how we talk about AI output and how that affects the way we justify our belief in the reliance or trustworthiness of that output.

To indicate, to affirm or to say something, is a predicate of things that can define what kind of thing we are referring to, simply because some of these predicates are distinct to different kinds of sources. Surely, "telling" or "saying" something is only instantiated by subjects, and whether or not we trust what we are told, is based on social or formal relationships. On the other hand, the indication of objective information is conventionally only attributed to objects like instruments. However, the ability to indicate truth is an ability that both subjects and objects can have in common. For instance, an AI-based weather app can indicate a chance of rain tomorrow, and a meteorologist can testify that there will be rain tomorrow, and both claims can act as indications. Yet, we trust the testimony of meteorologists and rely on the indications of an AI-based weather app.

A common justification for why AI-based beliefs are deemed reliable or "trustworthy" is based on factors like who made it who sells it and for what purpose it was made (like I will discuss further in section 2.3.1). But in general, I would argue that we, the general public, rely on the output of AI because it is known to be good at what it is doing, and in our experience, it is not often wrong. Or at least not wrong often enough to dismiss without consideration when using AI in the wild. In short, it is considered reliable because it is often reliable. Suffice to say, it is a rather circular justification, but also a reasonably pragmatic one that will be discussed in greater detail in section 2.2.1. What is important is that the theory of justification that we apply to believe different indications depend on how we refer to that indication. Some people may say that that the AI-based weather app "told" them that it was going to rain, but the implications of that are that we justify our belief in that app similarly to how we justify our belief in the testimony of a meteorologist.

1.3 AI as potentially many different sources of knowledge.

Because AI has such a wide range of epistemic features, it seems that AI as a general category can represent several kinds of external sources of knowledge that we justify our reliance on in different ways. In this section, I will introduce some of these problems and how they may lead to the need of placing AI in distinct epistemic categories depending on its epistemic features.

From an epistemological point of view, it's not always clear how to justify dependence on AI as AI is a rather broad term that includes a diverse selection of systems with different epistemic abilities, features and characteristics. Indeed, it seems AI systems can be quite distinct from each other, and possibly lack the usual transferability between them required if we want to make general epistemological judgements about AI. It seems that if we are going to compare AI with conventional sources of knowledge on a general basis, we must designate different kinds of AI into categories based on how their abilities are comparable with conventional sources.

What makes things even more difficult is if new iterations of AI has special features, characteristics or abilities that could contradict or negate previous expectations. It becomes clear that the range of systems that goes beneath the AI umbrella seems too large for any constructive analysis of specific AI beliefs as they might not have much in common. Furthermore, the technical configurations of AI are not always relevant to its epistemic status, as other abstract features, such as the degree of opacity and autonomy manifested by the entirety of the system may be more relevant. What becomes clear, however, is that we might need to separate different kinds of AI's into some basic classes based on their distinct epistemological abilities and limitations. An epistemological taxonomy of AI, if you will – that could be a helpful guide for reasonable justifications of AI-based beliefs.

1.3.1 introducing an epistemic taxonomy for AI.

Is it possible to make an epistemic taxonomy for AI? Here, I will argue why an epistemological taxonomy may be important for us to find the correct theory of justification for any given AI.

Talking about AI in terms of taxonomy is, perhaps surprisingly, quite common in the world of computer science. This sort of taxonomic perspective on AI is usually aligned with biology, where AI represents an ecosystem of different machines and programs. The distinct pieces of the AI are then sorted into different classes and categories based on the unique way each piece operates, essentially like different species, genus and kingdoms. This must certainly be quite useful for an overview of AI technology – but how well does this work for the epistemology of AI?

Interestingly, no such attempt has been made in terms of epistemological abilities that can account for the diversity found in different kinds of AI technology. However, dividing AI into different classes based on technology alone does not necessarily reflect the epistemological abilities of AI sensibly. It is rather how the particular configuration of each AI manifest abstract, epistemological abilities that are relevant for our theories of justification for our reliance, or trust, in the system as a whole. In this respect, the biological approach of AI taxonomy is excessive, as epistemology does not have anything synonymous with the structures and hierarchies of biological classifications. What we do have, however, is a set of sources of knowledge and justification, which are: perception, introspection, memory, reason, and testimony. With each of these sources follows questions about their reliability and trustworthiness, but they do not have an established hierarchy (family, genus, kingdom etc.) like the one we need for our epistemological purposes. Still, that is not to say that a simple way of organising AI compared to conventional sources of knowledge is impossible.

In the first part of this essay, I will argue that rather than technology alone, the opacity and autonomy of each AI model are the most relevant epistemological factors to consider when assessing the model's epistemic features. I reason this by claiming the abstract features of opacity, autonomy and complexity of *any* system fundamentally affects how we justify our reliance on it. As opacity or autonomy increase in a system, our ability to explain why we rely on it diminishes. Some systems are relatively opaque, but not very complex as such systems generally have very limited functionality, like an internet search engine. On the other hand, some systems have a relatively high degree of autonomy but are fairly transparent and simple – for instance, something along the lines of an autonomous vacuum cleaner. You also have systems that are opaque and autonomous, where our justifications change depending on the degree of

complexity. A honeybee can reliably lead you to a flower, but in a different way than Gary down the street does. Both a honeybee and Gary are fairly opaque as you can not explain exactly why or how they know about honey, and they both act on their own according to their agency (to some extent). You may find both reliable, but one is indicating where the honey is, almost like an instrument or perhaps as an epistemological agent (depending on who you ask), and Gary is testifying that he knows where the honey is.

In philosophy, we have distinct ways of justifying beliefs in testimony compared to other sources like observations of a honeybee or the readings of an instrument. The degree of autonomy and transparency of each system (i.e. object, organisms or people) defines the premises for why we should rely on each respective source in a way that is also relative to its degree of complexity.

To give an impression of how the basic structure of such an taxonomy would work [figure 1] (see appendix) illustrates a basic grid of opacity and autonomy. In this grid, sources can be divided into four different classes, based on a rough estimate of their positive or negative opacity and autonomy. This grid does not take into account the complexity of the system, but rather if it features a substantial degree of epistemic opacity and autonomy that affects their epistemic status. [Figure 2] takes into account the complexity of the system and gives an impressionistic overview of how these epistemic dimensions intersect. Depending on the truth value of the disjunction, we can make some basic assumptions about what kind of justification is appropriate for each combination. (See figure 1). This leads us to my suggestion of a tentative epistemological taxonomy of AI.

1.4 A tentative epistemological AI taxonomy.

In this section, I will introduce the AI HLEG's taxonomy of AI and discuss some of its implications, and how that leads to my tentative epistemic taxonomy.

Fortunately, the High-Level Expert Group can provide us with a proposed AI taxonomy that is, indeed, based on the capabilities of AI. Although these capabilities are not necessarily aligned

with epistemic capacities, we can use them as a starting framework for our discussion of the epistemological taxonomy of AI.

The distinct capabilities are, according to HLEG, **(i)** reasoning and decision making & **(ii)** learning and perception. (High-Level Expert Group on AI, 2019, p. 11)

(i) Includes what HLEG describes as "*the transformation of data into knowledge, by transforming real-world information into something understandable and usable by machines, and making decisions following an organised path of planning, solution searching and optimisation." (ibid.)*

This is the kind of AI we probably tend to interact the most with on a daily basis as it usually involves the kind of software that helps us with planning, organising, optimizing and searching the internet. It uses symbolic rules to make fairly transparent inferences to generate its output. In other words, the inferences this system makes is reducible to a symbolic ruleset that can be deconstructed or explained by reduction with reasonable effort. However, as the HLEG highlights, the process of information transformation is first and foremost to make it useable for machines – not people. This is important as it underlines how AI systems tend to become opaque.

Furthermore, note that HLEG use the word "knowledge" in a different way than the way we do in philosophy, and it is better if we rather think of it as "output" as for now - before we make any assumptions that AI has the capacity of possessing knowledge – or even being able to pass knowledge onto us.

(ii) Includes systems that have the capability to *learn* instead of depending on symbolic rules to make inferences. "Learning" is the key-word here; HLEG describes "learning" in this context as: *"meaning the extraction of information, and problem-solving based on structured or unstructured perceived data (written and oral language, image, sound, etc.)-, adaptation and reaction to changes, behavioural prediction, etc.".* This type of AI is typically associated with the fields of AI that include, but are not limited to; Machine learning, Natural language

processing, and Computer vision. The fact that this kind of AI has the capability of learning autonomously is distinct from most types of instrument-based sources of knowledge.

Let's take a moment to look at the epistemological difference between (i) and (ii). One important distinction between (i) and (ii) from an epistemological point of view, is that (i) operates in a potentially more transparent way than (ii), so justifications of belief in the reliability of (i) are easier to explain than with (ii). However, that is not to say that these types of AI necessarily operate separately from each other, or, in a way that makes it easy to distinguish if the output is a result of either (i) or (ii). Indeed, these technologies are often implemented in systems bilaterally, and thus the boundaries between the shared output are often fuzzy for both laymen and experts. This means that the epistemic status of AI does not necessarily depend on it being a type (i) or type (ii) AI, but rather how the combination of the two manifests epistemically opaque or epistemically autonomous features.

Still, the HLEG's proposed taxonomy is very valuable for us, as it highlights one very important epistemological difference; namely their epistemically autonomous and epistemically opaque features. In the next section, I will discuss some of the important epistemic implications of a system having epistemic autonomy, and in section 1.5, I will discuss the meaning and implications of epistemic opacity.

1.5 AI autonomy.

In this section, I will briefly discuss AI as a concept and how it is used in the context of AI. To begin with, it is worth noting that it is not uncommon in other fields to talk about autonomy in terms of operation, and not in the philosophical way that refers to different uses of the will on account of mental content. However, it seems that AI has the potential to gain autonomy in a philosophical sense, so both a strong and weak interpretation of autonomy will be discussed. Strong instances of AI autonomy will be discussed further in part two of the essay. Let us begin with a short discussion of the general use of autonomy in the context of AI.

In textbooks and research about AI (including the HLEG definition), the word *autonomy* is thrown around a lot without making it explicit about what they mean with this terminology in the context of AI. Although some may say that what is autonomous is self-evident, as the word is derived from the Greek word *autos*; meaning "self" and *nomos*; meaning "rule". Thus, if something governs itself (i.e. rules itself), then it is an autonomous thing. Unfortunately, this is not as straightforward as it might seem. The concept of autonomy has long been a contested topic in philosophy, with little to no consensus about its implications or even its plausibility. It is generally associated with problems concerning peoples free will how that is connected to different kinds of agency, but it has rarely been discussed in the context of real objects. There are some famous examples of thought experiments, like The Chinese Room Argument (CRA) (Searle J. R., 1980) that discuss the metaphysical challenges facing AI regarding "will" in terms of intentionality. However, these discussions have been centred around metaphysical questions of consciousness, and not the epistemological implications of AI consciousness. The notion of an object having any sort of *will* is an extremely complicated topic that is beyond the scope of this paper. Instead, I will discuss autonomy in terms of the implications that autonomous *behaviour*, have on the way we justify our belief in its reliance or trustworthiness. This is something we may see in current and near-future AI technology, and it is indeed a cause for concern for its stakeholders that depend on its reliability or accountability.

1.5.1 Introducing some epistemic implications of autonomous AI.

Why does AI autonomy matter for how we justify our belief in it? That depends on what we imply when we say autonomous. It is quite interesting that the field of AI generally uses this contested concept as a part of their dialectic without making it clear what exactly they are implying. Autonomy, in an epistemological context, will always be connected to agency to some extent. That is to say, the ability to act on its own will. This means that as the degree of autonomy increases, so does its agency and thus the reliability or trustworthiness of the output of AI with high degrees of autonomy depends less on how it is modelled – and more on how it uses its agency. In other words, it may have preferences, and its behaviour becomes harder to predict as we have to account for the possible preferences of the system. This element of "self-governing" must factor into how we justify our reliance on AI because a high degree of agency changes the premises for justification.

Let's consider some roundabout strategies that could plausibly make the problem less complicated. One strategy to avoid all of this would simply be to change the terminology to avoid the negative implications regarding any sort of actual agency. By rejecting the notion of truly autonomous behaviour in AI and writing it off as a mere imitation of actual autonomy. We can simply write it off as something like an illusion caused by programing that only imitate actual autonomous behaviour – similar to the CRA. Perhaps the term *"autarky"* is better suited then, as it does not involve self-governance, but rather self-sufficiency – which is a less contested concept that does not imply a will or intentionality to any extent.

Another term that may be more appropriate is "*automaton"*, which is more traditionally used in the context of robots and other automated machines that give the impression of acting on their own. "*Automaton*" is also derived from the Greek word "*autómaton*", meaning "self-moving", but this concept comes without the same difficulties as *"autonomy"*, as it does not extend beyond machines that can operate by themselves. However, since *automaton* has its distinct historic usage, perhaps *autarky* is better suited. Either way, some may say that any other term than "autonomy" is more appropriate when referring to any kind of object, as we do not face the same epistemic implications with those terms as we do with "autonomy".

Unfortunately, simply changing the terminology does not free us from the burden of explaining the behaviour of AI that can make its own input (self-governing, in a weak sense), as some AI can potentially do. After all, there must be a reason why computer scientists like the HLEG put so much weight on AI having autonomy. Perhaps the best way to try to work with this concept rather than denying it outright and rather try to understand how functional autonomy can affect its epistemological status.

1.5.2 Different types of autonomy.

In philosophy, we usually make the distinction between personal, moral and political autonomy as they all have very different meanings and implications. This calls into question, which one does HLEG have in mind when they use it in their definition? Furthermore, the type of autonomy – and the degree of authenticity of that autonomy – affects our justification of belief in the reliance, or trust, in AI. All three types could be relevant, but some may be more realistic than

others. In this section, I will briefly discuss these possibilities and show some examples of why this may cause issues for how we rely on AI.

For starters, it is unlikely that HLEG is referring to the potential political autonomy of AI as it is not plausible that AI will be involved in politics any time soon – although it sounds promising as an exciting sci-fi premise for the future. What I do believe HLEG has in mind when they refer to autonomy, is a sort of personal and\or moral autonomy. Keep in mind, that according to their very definition, autonomy is something AI must possess *"…some degree of…"* (High-Level Expert Group on AI, 2019, p. 1), but it seems like up to epistemologist to determine what this exactly means for our reliance, or trust, in AI. This is something that I believe deserves further philosophical attention and discussion. However, to keep this essay as grounded in current reality as possible, autonomy will be interpreted simply as "functional autonomy". In short, autonomy without intentionality or the mental states that are often associated with it.

While this question as a whole is far too complicated to sort out here, I will give a short explanation of why this matters for theories of justification of belief in the reliability, or trust, in AI. In the following section, I will give some examples of AI acting autonomously, and why that may be problematic for the accountability of, and thus the reliance on, autonomous AI behaviour.

1.5.2.1 Autonomous AI, authenticity, and accountability problems.

If we can say that AI is autonomous, can we say that AI can be authentically autonomous as well? In the philosophy of action, there is an ongoing debate about what "authentic" autonomy is, including what conditions must be satisfied for an action to come from the agent's "true self". In the context of AI, this might be transferable to questions about what output spontaneously originated in the AI, and what is a result of the epistemic features of the AI model. Such behaviour can align both with personal and moral autonomy, in the sense that it not only acts by itself as a moral agent but also makes other choices based on its personal preferences that are opaque to outside observers. These preferences may contain problematic biases, like racist or chauvinistic inclinations that are hard to anticipate for its designers. This has already been observed in AI without authentic autonomy that has been fed biased datasets either by accident, or faulty channels. In these cases, the AI's autonomy has led to it behaving in ways its designers

did not foresee or was able to explain with certainty. When this behaviour is problematic or causes harm, an accountability problem emerges, as it is difficult to hold anyone directly responsible for its behaviour. Especially in cases where this seems to happen spontaneously, it is difficult to find a theory of justification for the reliance on AI as an instrument that can take this into account. In fact, there are already instances where AI has been implemented in the wild, where it has caused unforeseen problems and accountability issues.

1.5.2.2 Examples of AI accountability problems following AI autonomy.

One example of this was Microsoft's infamous chatbot Tay the teenager, which was designed to use machine-learning modelling to learn about the world from the internet – in particular Twitter. Within a day of interacting with the internet, Tay had become an outspoken racist, misogynist and Holocaust denier, and was soon banned from Twitter. Mircosoft, embarrassed by the result, attempted a re-launch the bot days later resulting in Tay getting stuck in a loop re-tweeting *"You are too fast, please take a rest"* over and over again, aggravating its 200 000 subscribers with spam. After the public relations disaster of Tay, Microsoft stated that the lesson they learned the most is that they need to find a way to make AI more accountable before allowing it to interact unsupervised with the public. Here we can see that Microsoft encountered problems by believing that Tay could safely and reliably interact with the Twitter environment autonomously. Furthermore, while Microsoft took responsibility for Tay (Lee, 2016), they commented on the difficulties of designing an AI model that can be accountable for spontaneously bad behaviour.

Another, more example with more serious consequences, is an AI that was used by US courts as a support tool for risk assessments of recidivism, named Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). This system used machine learning algorithms to predict the risk of recidivism of criminal defendants. This AI was used by the prosecution in hundreds of courts across the US before investigative journalists revealed that it was almost twice as likely to deem African-American defendants almost twice as likely to re-offend (Angwin, Mattu, & Kirchner, 2016). Furthermore, some of the techniques used by COMPAS were protected as trade secrets, which hindered adequate transparency for both the defence of the defendants it was used against and public scrutiny. While there is no way of knowing for certain how accurate the output of COMPAS was, beyond counterfactual reasoning. However, the

accuracy of the system has later been deemed to be only slightly more precise than layman's predictions of recidivism. Interestingly, information about the defendant's race was not included in the datasets fed to the algorithm (COMPAS Case Study: Fairness of a Machine Learning Model, 2020). By all accounts, COMPAS became spontaneously racist, without any clear way of holding anyone accountable for its output.

Cases like this, however, can often be traced back to some fault in the datasets that skew the algorithms towards unfair biases. Autonomy, authentic or not, adds another level of complexity to problems like this because it allows for biases to spontaneously appear without a clear cause that can be traced to its input. Normally we could use arguments like “garbage in, garbage out”, to blame the biases on something wrong with its datasets or some other bad configuration within the model. But if the AI is authentically autonomous, there is currently no clear way of taking into account such spontaneous biases. Again, if we cannot predict the behaviour of autonomous AI, it poses problems of justifying our belief in the reliance on such systems.

At present levels of AI though, we can understand autonomy in the context of AI as an ability to have some level of epistemic agency, without the accountability that we expect from epistemic subjects. At its lowest levels of complexity, this is no different than the rudimentary actions of an automaton making inferences based on a well-defined decision tree. At its highest (current) levels, it is more like an artificial animal that understands its environment and makes rational decisions according to the goals it has set. Hypothetically, this can continue and escalate up to Strong AI, where the intellectual capabilities of the AI are functionally equal to humans. This AI has full, (possibly authentic) autonomy – and at this point, its deliverances seem to be more suited as testimony than simple, instrumental output. AI testimony causes some challenges for reliance, or trust, in AI that will be discussed in section 2.4.3. My goal here is to give an impression of why AI autonomy is relevant to an epistemic taxonomy of AI, as the degree of autonomy, opacity and complexity can severely affect the sorts of justifications we are entitled to make. [Figure 2] (See appendix) is meant as an impressionistic illustration of how much these features matter. We can then see that AI can be compared to a variety of other systems that we have very different theories of justification for.

I have now spent some time talking about the epistemic relevance of autonomy in external sources of knowledge for how we justify or reliance, or trust, on them. Let's move on to why opacity is also a relevant epistemic feature for theories of justification.

1.6 Epistemic opacity

Let us begin by asking, what is epistemic opacity? Paul Humphreys describes epistemic opacity as follows: *"Here a process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process."* (Humphreys, 2009, s. 618). An implication of this is that we can not directly determine the behaviour of an epistemically opaque AI. It has an impenetrable quality that hinders epistemic insight into the systems inner workings by outside agents. And on the other hand, if it were transparent, the inner workings of the system would be apparent and explainable by observing agents as it is operating. The only way of explaining the behaviour of epistemically opaque AI, is by some sort of counterfactual reasoning, *post hoc* hypothesis or *ex post facto* research like suggested by (Jalota, Trivedi, Maheshwari, Ngonga Ngomo, & Usbeck, 2020). The important takeaway from this is that we cannot explain the AI's behaviour before, or while, it is in operation.

A completely opaque AI is often referred to as a Black-Box AI. To get a better idea of what a black box is, consider this: Imagine a Black-Box that prints receipts with information that seems accurate, but the process that caused the content of the receipt seems to be hidden behind the walls of the Black-Box. We may be tempted to break the box open to take a look at its inner structures, but when we do so, we can not find anything that clearly explains its behaviour. This may be because the structures we observe does not make any sense to us – that they are too dense, too complicated or seemingly random patterns that make it all seem unintelligible. But in the case of Black-Box AI – we designed it, so why can't we explain how it works? One of the reasons for the opacity is often caused by the sheer size of the datasets it is processing. It is practically impossible for scientists to determine how, for instance, petabytes (PB) worth of information lead to a specific behaviour. A petabyte is simply too much data for humans to comprehend or sort through manually in any realistic timeframe. Comparably, a Google AI

sorted through 1 PB in six hours and 27 minutes over ten years ago (Czajkowski, 2008). That said, it is not always the sheer size of the content that makes the system so impenetrable. In the next section, I will give a short explanation of why some AI technology such epistemically opaque features.

1.6.1 What makes some AI so opaque?

If we want to have a plausible theory of justification for belief in the reliance, or trust, in AI, it is helpful to understand why some AI has such a high degree of epistemic opacity. This is relevant in the sense that it highlights how epistemic autonomy and epistemic opacity are often connected – and why these features are important for the framework I suggest for a tentative epistemic taxonomy.

In recent years, one of the most powerful and productive AI classes has been the product of Deep Neural Network (DNN) modelling. It has opened up new technologies like computer vision for automated driving and new protein-folding configurations for biology and medicine. But what is it that makes DNN tach so intrinsically opaque? While there are many different DNN models, like Convolutional Neural Networks (often referred to as ConvNets) and Recurring Neural Networks (RNN's); they all function with high degrees of opacity. As the name has it, deep neural networks may seem like they are designed in a way that is supposed to mirror the way neurons function in a brain. Although the brain has certainly inspired deep neural network architecture and some computational neuroscientists use DNN's for their research (Savage, 2019); the analogy between DNN's and brains has its limits. Keep in mind, the "neural" in DNN does not refer to actual neurons. Instead, DNN models have a deep stack of layers that are hidden beneath an initial input layer before reaching the bottom output layer. This is what makes DNN's "deep". Each of these layers consists of non-linear module representations that are refined as the stack becomes deeper. The model does this refinement by having an objective function that measures the degree of error by giving the output a score and the desired pattern a comparable score. As the model reduces these errors it adjusts its internal parameters to mitigate its errors. These parameters are often referred to as "weights" that are adjustable for each layer. The opacity issue is then caused by that in a DNN there can be hundreds of millions of these non-linear weights and a hundred million more representational examples with labels connected to

these weights. This shows how AI's autonomy can lead to such high degrees of epistemic opacity, as highlighted by this definition of DNN: "*The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure"* (Lecun, Bengio, & Hinton, 2015). To make matters worse, this kind of Black-Box AI is also unable to explain its behaviour in a way understandable to us as it does not (yet) have the adequate self-awareness necessary for this action to be done in a reliable or intangible way. How then, are we supposed to justify reliance on opaque AI?

1.6.2 How opacity matters for epistemological justification.

Opaque sources of information can, in some cases, pose problems for how we justify belief in their reliability or trustworthiness. In most cases it is possible, at least theoretically, to explain how the sources we rely on are reliable. Objects that we rely on may be epistemically opaque to their users, but the people who designed the object can explain how it works with great certainty. These designers work in a society that holds them, or the industry responsible for how these objects are manufactured and marketed so that we can feel entitled to a reliable level of quality. In other cases, where we for instance pay for the service of someone, we know that their business often relies on the quality of their service and trustworthiness. We can say that there are societal norms and official rules and regulations in place that should ensure that we are justified in our belief that these products or services work. Next to all of this, we may also have experience with the product or the service, which assures us that these things are reliable. These reasons can all act as a part of our theory of justification for reliance or trust in the relevant product, service or testimony.

However, when we try to apply this to AI, we encounter some new issues. (1) The epistemic opacity of AI does not only concern its users but also its designers. As the designers have little, to no epistemic access to the inner workings of the AI, and thus they are incapable of assuring us of exactly why or how it will work. (2) When those who designed the system do not understand why, or how, it behaves – the expectations we are entitled to from the product in terms of reliability or trustworthiness is greatly diminished. As the AI is both autonomous and opaque, it is difficult to hold anyone accountable for the AI's behaviour or agency. (3) There are few societal norms developed context of AI and the industry that surrounds it. The standards for what

we are entitled to expect of AI in terms of quality, reliability and trustworthiness are still under development and the industry faces little regulation. (4) Our experiences with AI is not always sufficient for justifications for the future performance of the AI for two reasons. First, AI software is constantly changing in terms of new updates and new iterations. Because of the epistemic opacity of AI, both the designers and users are not always aware of how these changes affect the reliability or trustworthiness of the AI in question. Second, the opacity of the AI makes it difficult for its user to check if it functioning correctly. If, for instance, the AI solves extremely complicated calculations beyond the capabilities of its users, its users will have serious difficulties in checking if these calculations are correct, and also explaining how the AI arrived at its results. In cases like this, the AI may very well be malfunctioning with its users aware of this.

While not all AI is equally opaque or autonomous, we can see that these epistemic features can lead to some problems in justifying our belief or trust in it. In the next part of this essay, I discuss if these features are also found in conventional, external sources of knowledge, and discuss if some of the theories of justification we use for them, can also be applied to AI with similar features. While not all theories are discussed, I will focus mainly on the social justifications we use for reliance and trust on those sources and see how they compare to AI.

2 AI compared to conventional sources

In this part of the essay, I will discuss how AI can compare to conventional sources of knowledge and how the theories of justification we use to believe in their reliance or trust, can plausibly be applied to AI. As shown in [figure 1] (see appendix), AI can be divided into four classes (i.e. the four combinations of negative or positive opacity and autonomy) that may require different theories of justification.

It is worth noting that since the HLEG definition states that AI must be autonomous, some might say that we can only apply theories of justification for belief in trust and reliance that concerns autonomous things. I will argue that this is not necessarily true because simple AI does not have a significant degree of epistemic autonomy. It is better compared to simple instruments in terms

of how we justify reliance on it. In other words, low levels of epistemic autonomy or epistemic opacity are not always relevant for proper justification of belief in reliability. Furthermore, for all classes of AI, I will highlight that the social reasons we use to justify reliance or trust in AI are often the most important.

2.1 Four categories based on opacity and autonomy

Here I suggest that AI can be divided into four classes based on a rough estimate of their negative or positive opacity. By seeing how these epistemic features intersect on a grid of [figure 1], we can say that there are roughly four classes of AI that have distinct epistemic features. By using this as a guide to compare AI with conventional sources, I will in later sections discuss how we justify reliance or trust on those sources, and if we can apply the same justifications to AI.

Please keep in mind that few things are absolute in terms of opacity and autonomy and that this is only to illustrate how it is possible to understand AI epistemically based on these features. The purpose of this is to have a general, tentative taxonomic guide to the closeness AI can have to various sources as [figure 2] (see appendix) also impressionistically illustrates.

In other words, it is not meant as a conclusive characterization of each source, but rather a way of placing epistemically fuzzy sources like AI (in the sense that it has the potential to be similar to many things) among conventional sources. Here are some examples of external sources of knowledge that can have similar epistemic features.

2.1.1. Members of the four categories.

- [Opaque and autonomous]: This category includes rational agents, epistemic subjects, and epistemic agents. Typically people and animals. They act on beliefs, or functional beliefs, that is epistemically opaque.

- [Not opaque and autonomous]: This category includes different types of automation, like for instance robots on an assembly line. Their behaviour may be considered autonomous,

but with very limited epistemic autonomy. Their behaviour is predictable and can be explained by experts.

- [Opaque and not autonomous]: This category includes digital electronics that use fairly complex algorithms, but have a high reliance on manual input. While not opaque to experts, their inner workings are opaque to their operators.

- [Not opaque and autonomous]: This category includes analogue instruments and simple digital instruments. It is not always completely transparent to their operators, but they are not difficult to explain in theory. The theories of justification for reliance on this class overlap with conventional instruments and will not be discussed at length.

2.2 Justifications of sources of knowledge that are neither opaque nor autonomous [-O.-A].

This section starts with referring to theories of justification that may be sufficient for sources of knowledge that have the [-O.-A] epistemic features. What is, for instance, a common theory of justification for belief for instruments or objects we believe in the reliance of on in the wild? Let us first consider what a common-sense theory of justification of reliance that most people apply, could look like.

2.2.1 A common-sense theory of justification for belief in reliability.

In general, we justify our belief in reliability in everyday objects and instruments for common-sense reasons. This common-sense theory is rarely made explicit, yet I suppose these are the most ordinary justifications for a general belief in the reliability of instrumentation we make in everyday situations. As reliance on AI is becoming increasingly instrumental in many peoples lives, this theory of justification may apply similarly. At its essence, it is a reliabilist theory of justification, that may apply to more than just [-O.-A] sources of belief, but since it is the most basic, common-sense theory, let us first discuss it in the context of simple instruments. Let's call it The Simple Epistemic Theory of Instruments (SETI), and let us first see how it applies to conventional instruments that do not include AI.

Here is my tentative definition of this theory: We find beliefs in the reliability of instruments to be justified because; (A) instruments are known as reliable objects, or (B), the user's experiences is that the belief in the reliability of instruments are usually successful beliefs. By "successful beliefs" I mean beliefs that turn out to be true.

Let us stop for a moment and consider what the two parts of this argument can imply, beginning with (A). Part (A) can perhaps be criticized as being somewhat circular if they suggest "Instruments are reliable because they are reliable". But is this criticism really valid?

(A) Is perhaps not a bad justification if we suppose that instruments are deemed reliable *prima facie* because being an instrument comes with a certain epistemic status. It also follows that this status should not be coincidental. Most people have a notion that the designers of the instrument are part of an industry that usually has a degree of regulations, standards and practices that entitles us to certain expectations. We might also know if the instrument has a generally good or bad reputation. Perhaps a friend has assured us through their testimony that you can believe in its reliability. We might have experiences with the brand or industry that makes them that enhances our belief that the instrument is reliable. Furthermore, if the instrument was sufficiently unreliable, it would not be able to contend with other brands or designs. There is a general notion that instruments must qualify for their status in a way that should guarantee a minimum of reliability. If we assume that all of these notions are a part of the reason it is referred to as an instrument, we might be justified to infer *a priori* that they are supposed to make reliable indications.

Looking at it this way, (A) can imply valid reasons to believe in the reliability of instruments. Although it might be accused of some degree of epistemic circularity by sceptics – most people seem to find (A) and the justifications it may imply, more than sufficient for belief in everyday reliance on instruments.

Part (B) of SETI can in many situations be a sufficient reason by itself. If someone has experienced how reliable, say, a thermometer is at telling the temperature, they can justifiably believe that thermometers are reliable indicators of temperature. The amount of positive or negative experiences the user has with the instrument create memories of good or bad reliability.

That is to say, if the belief in reliability are often successful, most people would feel justified to have those beliefs.

When we then take (A) and (B) together and consider all of the reasons they may imply, the SETI can be a very reasonable theory of justifications of instrumental beliefs in our daily lives. SETI can thus represent the set a priori and a posteriori reasons we often implictly apply as justifications for reliance on instruments. Some may say that the SETI is inadequate for scientific purposes or situations where the stakes are especially high, as the reasons that are given are arguably contingent. However, this level of scepticism is hardly practical in our everyday business. In the next section, we shall see if the same reasons SETI is a reasonable justification for reliance on instruments in the wild, can also plausibly be applied to AI.

2.2.2. The SETI and AI.

Is The Simple Epistemic Theory of Instruments sufficient for justifying beliefs in the output of AI? AI has little in common with analogue instruments as they are not digital and they often depend on a constant causal relation with its input over time. However, the may often be the way we justify our belief in AI systems in our everyday business. For instance, the AI apps that are implemented in many of the devices that are casually incorporated in many aspects of our everyday lives, like our phones, TV and car. In systems like this, the SETI seems plausibly sufficient, as experiences of reliability are enough for most people to justify their belief in simple systems that operate in relatively low stake situations. However, because of the epistemic opacity of AI, the reasons (1), (2), (3) & (4), given in section 1.5.2, prevents the SETI from being an adequate theory of justification for epistemological reliability of AI in more serious situations.

2.3 Justifications of sources of knowledge that have saliently opaque features, but are not autonomous [O. -A].

In this section, I will discuss Sanford Goldberg's theory of justification for reliance on instruments, named Epistemically Engineered Environments (Goldberg, 2017). I will argue in the following sections that this theory is well suited for justifications of reliance on opaque and not autonomous instruments. In short, this is a theory where the justifications for belief in the

reliability of instruments are found in the social dynamics between epistemic subjects and epistemic environments. I will argue this theory is well suited for beliefs in the reliability of many kinds of opaque and transparent AI that do not have a very high degree of autonomy. This includes the [-O.A] category, wich incudes instruments and robots are transparent and automated, but not highly autonomous.

2.3.1. Sandy Goldberg's Epistemically Engineered Environments.

Goldberg argues in his paper "Epistemically Engineered Environments", beliefs based on instruments are fundamentally justified on social norms, practices and standards found in an epistemic community that entitles us to what he calls "Design Dependence" (DD) (Goldberg, 2017, p. 2786). He also emphasizes the importance of what he refers to as epistemically engineered environments (EEE), which are the environments we learn to apply, trust and internalize the norms, practices and standards that give us reasons to rely on instruments. EEE's are often found in academic institutions that are arranged in such a way that it is unlikely that epistemic subjects will be exposed to unreliable information. In a sense, it is an environment that can reliably provide knowledge in an easy way for the people participating in it.

Can these justifications plausibly be applied to AI as well? First, let us take a closer look at his arguments.

Goldberg also argues that one of the most important reasons for believing in the reliability of instruments is that instruments are designed, manufactured, tested, marketed and sold by epistemic subjects that are held to certain expectations. In turn, this creates a certain social dynamic that creates reasonable expectations of the industry behind the product. The expectation is that they must be able to guarantee a level of accuracy and reliability to live up to their epistemic responsibilities. Goldberg refers to this in short as "Design-Dependence" (designers referring to the industry as a whole) and it reflects how society as a large has developed an epistemic dynamic with designers of instruments that very much depends on the designer's epistemic responsibilities. The role of the designers, manufacturers and the industry seems to play an undeniable part in the epistemic status of any given instrument.

These reasons are not unlike the reasons I argued for in section 1.5.2. And indeed, DD gives many good reasons for justified belief in the reliance on AI. However, Goldberg highlights not only how these social practices lead to knowledge of epistemic norms in the epistemic community, and not only beliefs. That is to say, if we have knowledge of epistemic practices, we also have much stronger justification for the reliability of those practices than when we have mere beliefs about them. Goldberg argues that this knowledge, in addition to what we know about the marketing and education we have about the instruments that we buy, gives rise to justified expectations of reliability. If those expectations are not met, then society is structured in such a way that we are entitled to compensation, which in turn acts as a deterrent for the industry behind the instrument to make unreliable instruments. In a sense, this is based on a system of accountability. However, like discussed in 1.4.2.2. accountability can be a serious problem for highly autonomous AI in a way that Goldberg fails to account for with "design dependence". As argued earlier, social norms, practices and standards are underdeveloped in the context of AI as it is still a fairly source of knowledge. Furthermore, the industry behind AI does is not yet sufficiently regulated in a way that entitles us to expect reliability or trust. Perhaps the solution is found in epistemically engineered environments?

2.3.2 Golberg's epistemically engineered environments.

What Goldberg coins as epistemically engineered environments, are social constructions that are designed to lighten the epistemic burden of a single individual. Just like lessons in a classroom, Goldberg argues, instruments are so well implemented in the structure of society that we are entitled to rely on them. In the future, when AI will most likely become a part of these epistemically engineered environments, perhaps then we will have developed a more epistemically sophisticated relationship with AI, in the sense that the EEE is more familiar with its epistemic limitations. Perhaps then we can rely on EEE as a justification for our belief in the reliability of AI. At this point though, this does not seem to be the case, as even the biggest tech companies in the world, like Microsoft, seem to make mistaken expectations of AI as seen with "Tay" in section 1.4.2.2.

2.3.3 Problems with Design Dependence justifications for more advanced AI.

Unfortunately, once we try to apply Goldberg's analysis to more advanced, more autonomous types of AI, the arguments fail to bring the same level of justification as it does with simpler AI. In comparison to Goldberg's arguments for "design dependence", highly complex AI with strong epistemically opaque and autonomous features is quite different from the kind of instrument his theory of justification is focused on.

One of the reasons is the fact that some AI largely designs itself. The designers make the models foundation, but in cases as DNN, the most effective models trains themselves unsupervised by any designers. Some are not designed for a single purpose and because this type of AI is still new, there has not been engineered any epistemic environments that are developed and sophisticated enough to support a sound theory of justification. Another important key difference is that a sufficiently advanced AI often has a level of autonomy that a "simple" system like a conventional instrument has. If we assume that AI can become completely autonomous, its epistemic features seem to speak against "design dependence". That is to say, the designers can make the AI model, but not explain or necessarily predict its behaviour, so any marketing or promises that the designers make on of the AI would have to be quite tentative and unspecific compared to conventional instruments. In fact, some consider AI to be rational agents rather than conventional instruments. In the next section, I will discuss this kind of AI and how it may cause issues for some of the social justifications found in DD.

2.4 AI that is both opaque and autonomous. [A.O].

This section is regards systems that act as rational agents on account of their opacity and autonomy. Some may ask why members of this class can be considered rational agents. To answer this, let us first ask; what is a rational agent? In the next section, I will give a short explanation of some of the most prominent theories of a rational agent. In the subsequent sections I argue that in some cases, it is possible to consider AI as a rational agent.

2.4.1 Rational agents.

In a broad sense, a rational agent is simply a subject or an entity that acts rationally to achieve its goals. It does so by applying its intellectual capacity to determine the best course of action to achieve what it desires. However, this broad understanding of rational agents can seem to include AI. In fact, AI has often been referred to as a kind of rational agent in computer science. In perhaps the most prominent textbook on AI, "Artificial Intelligence A Modern Approach" (Russell & Norvig, 2009) AI is indeed defined a rational agent as something that *"A rational agent is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome.*" (Russell & Norvig, 2009, p. 4). This describes what is known *as perfect rationality* – a concept that is most likely unrealistic as it postulates something that *always* does the right thing. As discussed in regard to the HLEG definition, using terms like "best" may be problematic as it is not only hard to define what is "best", but also because knowledge about the "best" expected outcome seems like an unrealistic normative claim.

However, Russell and Norvig concede that this, indeed, is unrealistic, but only because the computational demands are too high (Russell & Norvig, 2009, p. 5). I think a descriptive approach is more constructive in this context. I believe it is fair to say that an intelligent entity that uses its intelligence to make rational decisions in accordance with the agents desired (or set) goal, describes a rational agent sufficiently. Either way, both definitions apply to humans – and possibly some kinds of AI and animals. But how does AI compare as rational agent? This will be discussed in the next section.

2.4.2 AI as a rational agent

While it may be possible that some kinds of AI can, indeed, be considered a rational agent, should this not affect the epistemic status of AI as well? If we follow the grid in [Figure 2], we can see that highly complex, highly opaque, and highly autonomous AI at some point comes closer to animals and even subjects as it moves further away from instruments. Although it is hard to pinpoint exactly where this line is drawn, at some point, an AI changes from an instrument to a rational agent. But what kind of rational agent is this? Bringsjord & Govindarajulu (Artificial Intelligence, 2020),claim that the the goal of AI technology is to make

artificial animals. Some say, ultimately, the goal is to make artificial people. In the next section, I will discuss this possibility and how it affects our justifications for the belief in the reliability of AI.

2.4.2.1 Artificial beings.

When AI functions closer to a subject rather than an object, we are facing the potential emergence of an artificial being. This concept is often referred to as strong AI. Strong AI is, simply said, the notion of AI with consciousness. However, while we are currently a long way technologically from making AI that can be realistically be considered on par with animals, the notion of strong AI is popularly accepted in the technological community. However, the very concept of strong AI is seen as unsound by many philosophers. Searle has famously contested the validity of this concept with his thought experiment "The Chinese Room Argument" (CRA) (Searle J. R., 2014), which has proven to be hard to defeat and is rather ignored or outright denied by AI practitioners as mere philosophical fancy, without offering substantial rebuttals. While this debate is worth a mention, it is outside the scope of this paper, as the focus is on the epistemological status of current – or near future AI.

If we follow the definition of a rational agent as given in 2.4.1, there is a claim that AI, is *functionally* a rational agent. Furthermore, if we go by the tentative epistemological taxonomy I propose, the epistemological status of AI can be considered much closer to that of a rational agent on par with an animal, as such. Keep in mind that this does not postulate consciousness in AI, rather that it *functions* epistemologically as things with consciousness. So, if the epistemological status of AI is considerably close to an animal and a rational agent, what kind of justification do we have for belief in animals? Typically, when an animal acts as a source for belief, we choose to relate to that as either a source of instrumental knowledge or a source of testimonial knowledge.

Let us first consider how we treat the behaviour of animals as instruments with rational agency before we consider an animal that can give testimony.

2.4.2.2 AI as a rational animal agent and instrument.

In this section, I will discuss how instruments with rational agency affects how we rely on that instrument, and in turn how that matters for a theory of justification for those instruments. I will argue that ethical concerns will be the primary reason why design dependence type justifications can be inadequate for implementations of this kind of AI – particularly when reliability is a matter of safety. As we will see, the epistemic status of being a rational agent weakens the reliability of the AI, rather than strengthens it as we are faced with a responsibility gap.

Let us begin with how animals can be considered rational agents and how we justify our reliance on them as instruments. One example could for instance be a police dog that smells drugs on a suspect. In those cases, the dog is treated as an instrument by the police, and the marking of the dog is seen as an indication – and a justification for the belief that the suspect is carrying drugs. But again, the dog is only considered a rational agent in the sense that it has the rationality required to follow its training. It is still treated as an instrument. If we consider those who trained the dog the designers of the dog, in the sense that they made it into a drug-sniffing dog, we can still justifiably rely on the dog as an instrument. Does this mean our belief in the dog as a rational agent justified the same as it is with an instrument? Or in general, does an instrument with the capacity of rational agency affect our justification for belief in it? I would argue that there is an important difference in how we hold a dog accountable and how er hold an instrument accountable. The makers of the instrument will always be held accountable if there is something technically wrong with it, while the dog is often held to a different standard. For instance, if the dog bites someone, the handler of the dog is usually held accountable for the dog's actions, similar to how the makers of an instrument are.

However, the dog is sometimes also blamed to some extent if it behaves in a manner deemed too unreasonable to hold its trainer accountable for. In some cases, if the dog failed to follow its training and caused great harm even if the trainer did everything correctly, it would seem strange to write it off as simply a malfunctioning instrument. Malfunctioning instruments have a mechanical or technical cause that, in the end, it would be possible to blame the malfunction directly on. But in the case of trained animals, the malfunction can not be traced back to anything more than the animal’s agency and the animal is held responsible to some extent. These are cases

where we have reasons to believe that the animal intentionally did something bad. Sadly, in some severe cases, the dog is deemed a "bad dog" and will be euthanized if it has shown severe aggression against humans. Of course, this is not to punish the dog, but because it is deemed to be an unreliable and potentially dangerous agent. In a sense, it is held responsible for its agency. On the other hand, a malfunctioning instrument may be destroyed, but the responsibility is always put on its designers, manufacturers, or its operator. This shows that there is a gap of responsibly that occurs when the function of an instrument is found in the instrument's rational agency. When we justify our belief in the reliability of a rational agent, the agent may be held responsible for the way it acts that cannot be reduced or traced back to, someone else.

This is where we are today with some AI operated machines, in the sense that although the AI is autonomously operating the machine, it requires supervision because putting ethical blame on objects is irrational – even for objects considered a rational agent. In case the AI makes a mistake that causes damage, the designer, manufacturer, or operator should be accountable for the damage. This is why the driver of an autonomous vehicle is required by law to have their hands on the steering wheel and pay attention at all times because the software is incapable of being held responsible for its action. The idea of destroying a car that got into an accident on account of its automated driving system, because it is a "bad car" seems absurd. Even if the car is considered a rational agent, the car cannot mean what it does – it is simply following a program. It is simply incapable of having bad intentions. Instead, we would either blame the manufacturer, designer for selling a "lemon" or overpromising the car's capabilities – or we would blame the driver for not interfering when the system was obviously misbehaving. What follows, is that the reason we feel justified in the belief that the AI is reliable as an instrument is that there is an environment of epistemic subjects that can be held accountable for the behaviour of the AI. This aligns largely with a design-dependence theory of justification, but while a consequence of DD is that we are entitled to completely rely on some instruments, the same does not go for AI function as rational agents.

We can see that a gap emerges between the designer's intentions and the final products epistemic features and behaviour. This is similar to how Matthias (Matthias, 2004) and Sparrow (Sparrow, 2007) argues for the emergence of a gap of responsibility between the designers and the final

product. Ethical concerns restrict our entitlement to complete reliance on AI, such as AI drivers, because the AI cannot be held responsible for its own rational agency and the designers lack the ability to predict the AI's behaviour. In these situations, we need, at a minimum, an operator that can supervise the behaviour of AI, and intervene, in case it behaves badly. Let us call this soft reliabilism. It will require a reductionist approach similar to Design Dependence, but while Goldberg is satisfied with epistemic subjects in terms of designers and the epistemically engineered environments that surround the use of the instrument, this also requires the direct supervision of a competent person. It particularly applies to high stakes situations, where we can not responsibly rely completely on the rational agency of the AI and we must, at a minimum, have an operator that is accountable for its behaviour. We seem to be entitled to rely on AI to a degree, but less so than we are with AI that is not considered a rational agent.

2.5 AI as a rational agent that can give testimony.

In this section, I will discuss the possibility of highly complex AI that can be considered epistemic subjects that give testimony rather than output. I will argue the testimony of such AI is not as reliable as a human expert as its testimony lack the semantic content we expect from conventional testimony. What follows, is that the social reasons we have to justify our belief in the other sources that have been discussed fail in the context of AI testimony, as it is not capable of being social. I will also argue later, in section 2.5.3, that while AI may be a rational agent, it cannot act as of an epistemically engineered environment as an epistemic agent. What follows, is that we can not rely on AI testimony the same way we rely on human testimony in both reductionist and anti-reductionist theories of the philosophy of testimony.

Let us begin by asking, what is required for an AI to give testimony? If AI can give testimony it implies that it has a belief that it wants to communicate. There are two interesting metaphysical features of this: the AI "believes" something, and it "wants" something. To understand why this matter for the epistemology of AI testimony, we must first try to understand what they mean in the context of AI.

2.5.1 AI consciousness.

To want something or believe something, implies consciousness. Whether or not this includes semantic content or intentionality like mentioned earlier in section 1.4, this is a metaphysical problem beyond the scope of this essay – but for the sake of this essay, let us ground AI firmly in current technology and assume that AI cannot have semantic content (meaning). Furthermore, AI does not have phenomenological consciousness like living subjects have. Phenomenological consciousness is, in short, the experience of being a subject, like you experience being you – from the inside. This means it cannot subjectively feel pain or emotions, and it does not have the capacity to value social feelings like compassion, respect, and remorse. To it, these are simply syntactical values devoid of semantic content. However, it is conceivable that the AI model has an architecture so similar to how our cognition functions, that it can imitate every cognitive feature of our brain. That is to say, the architecture of the system can compensate for all the features of phenomenological consciousness, but without the actual phenomenological states that follow. Susan Schneider (Schneider, 2019, s. 49) has described this as cognitive consciousness, or functional consciousness. It functions as a subject, but it cannot mean what it says. It simply imitates semantic capacity in a way indistinguishable to us.

What matters, epistemically speaking, is that it functions in a way we can only compare to other epistemic subjects, and so, it might seem like it is epistemically equivalent as well. But as we shall see, a testimony from a functionally conscious epistemic subject has less epistemic worth than what we expect from people, and especially epistemic *agents,* because it lacks the social capacity that justifies our trust in them.

2.5.2 Is AI testimony less transparent than human testimony?

In this section, I will talk about if AI reaches a certain level of opacity, autonomy, and complexity, conventional reliabilist and reductionist theories of justification for belief in testimony is implausible for that AI. Let us take a moment to consider how it compares to the testimony of people, and even expert testimony.

Technically, this AI is an epistemic Black-Box, but so is the human mind. Just as we cannot explain the behaviour of this AI, no known theory of the mind can account for all human behaviour or how the mind works. In this sense, both human testimony and AI testimony could have similar degrees of opacity. Currently, the human mind is arguably the most epistemologically opaque system known to man, other than perhaps a gravity singularity such as a black hole – AI may develop similar levels of opacity and become what is known as a technological singularity.

In contrast to the human mind, AI models are currently made by humans, so counterfactual reasoning about its technical properties may be plausible. If given enough time to examine its testimony and how it correlates with its architecture we could plausibly explain certain behaviour *post hoc*. But what if the AI that gives testimony is created by another AI? Let us call this instance of AI; AI*. In those cases, the model created by an AI may be even more intelligent than its parent AI for reasons that human-AI experts cannot explain. Some, like Nick Bostrom (Bostrom, 2017), talks about this as a superintelligence, where the AI's intelligence grows exponentially as it learns to create better iterations of itself. This could have serious implications for how we justify our trust in AI*. If AI reaches superintelligence, it may be impossible to understand why AI* behaves the way it does.

Even if a superintelligence does not occur, an AI made by another AI may cause enough problems for any theory of justification for belief in its testimony. In those cases, the AI* may be considered just as opaque as a human, or even more so. This is because we have some knowledge of the human mind and can apply psychological theories to plausibly explain human behaviour to some extent. Attempting to do the same with AI*, would be anthropomorphizing a system that we have cannot justify any similarities to. Perhaps some would say that psychological analysis can plausibly be made for an AI that has a cognitive architecture like us. But doing the same for AI* would be unplausible because it is based on technology completely unfamiliar to us. Any level of anthropomorphising AI* testimony that cannot plausibly be justified similarly to how we can with AI testimony. The epistemological implications of this are that it is difficult, if not impossible, to have a reductionist account for AI testimony. It seems we must either choose to trust it or not. Let us now move on to the notion of AI trustworthiness.

2.5.2.1. AI trustworthiness.

As mentioned earlier, the AI HLEG's notion of trustworthiness may not be suitable for most instances of AI, unless it is in the context of AI testimony. The question that follows from this is, should we have a reliabilist theory of justification for the say-so of these entities? Or should we look elsewhere in the philosophy of testimony for a justification of belief based on this? Let us take a moment to discuss the possibility of untrustworthy AI.

The philosophy of testimony has a long history of discussing the epistemological implications of human testimony, but so far instances of testimony from AI has not been discussed much. However, the idea of AI as a source of trust or distrust has been discussed to some extent, which is closely connected to reliance on testimony. Billy Wheeler (Wheeler, 2020) argues that AI (or robots as he refers) is capable of both testimony and trustworthiness because it can be programmed to deceive us. C. Thi Nguyen (Trust as an unquestioning attitude, forthcoming) argues that objects can, indeed, betray us as we internalize our expectations of objects by having an unquestioning attitude towards them while we use them as a part of our agency. He argues that this is similar to how we, at times, can feel betrayed by our bodies and cognition. This notion of deception and betrayal is closely connected to the notion of AI trustworthiness, as we also have seen brought up by the AI HLEG (2019). This was a notion I criticized as I believe reliance is a more appropriate term for objects and the kind of AI we see today. But if AI advances to the point where it is capable of testimony, the trustworthiness of AI is crucial for how we justify our belief in it.

This leads us to two questions; (1) if we assume that AI is capable of not only testimony but also deception and betrayal, can we plausibly apply a reductionist theory of justification for its testimony? (2) If we assume that AI is capable of testimony, but not capable of betrayal or deceit, can we plausibly apply an anti-reductionist theory of justification for its testimony? In sections the next sections I will discuss the plausibility of theories of justification for such instances.

2.5.2.2. Arguments for trustworthiness.

Some may say that AI trustworthiness is not a problem for two reasons. The first is that they are shown to be statistically reliable. The second is that they have less reason to deceive than people.

The first reason would be that although the AI has functional consciousness – Design Dependence is still a sound justification because the AI has been tested so thoroughly by its designers that the probability of it being incorrect is highly unlikely. Their evidence simply shows AI's testimony is very reliable. More so, the data might show that the AI is statistically more reliable than human experts. In this sense, the opinions of the AI should be considered as good, if not greater than the testimony of human experts in terms of reliability and expertise. They may show that the AI is able to explain itself to a great extent and provide evidence that its findings are correct. And although we are unable to explain how it does this, all of these reasons speak to why we should trust it as much as any human expert.

The second reason is that they have less reason to deceive. Indeed, some proponents of this AI would likely say that it is no longer a question of reliance, but a question of trust. And if we can trust human experts, then we can trust this AI as an epistemic subject just as much as anyone else. After all, it is not as capable of having bad intentions as humans can. If the AI cannot have intentions and its designers have tested its reliability time and time again, we should welcome it to join our fraternity of experts as an equal. In fact, there are so many instances of human experts acting in bad faith that AI can be trusted even more, as it is incapable of having genuine intentions. Without intentionality, any motivation for deceit should be less common. Because of this, AI acting in bad faith would seem less likely and its trustworthiness is even greater than what we can expect from humans. According to this line of reasoning, AI testimony can be relied upon and trusted, and a theory of justification for belief should be anti-reductionist and reliabilist.

But does this line of reasoning hold up to criticism? In the next section, I will argue that AI lacks the social capabilities that are needed for trust-based justifications. Furthermore, by giving AI the status of an epistemic subject, it becomes part of the epistemically engineered environments in a way its epistemic features do not support.

2.5.2.3. Arguments against trustworthiness.

In this essay, we have seen that social reasons have been an important part of how we justify our belief in the reliability of each class of AI. Design Dependence, epistemologically engineered environments, and the testimony of people, consist of epistemic subjects that we assume take their epistemological responsibilities seriously. This capacity, however, is something AI is unlikely to have. The fundamental problem of both trust and reliance on AI testimony is that it is capable of deceit, while not having the capacity to mean what it says. In a sense, the words of AI will always be empty, as long as it is not capable of semantic content. What follows from this is that AI can have the negative epistemic features of testimony such as deceit, but not the positive, social features that we often rely on.

It is capable of deceit because it can have desires or beliefs in a functional sense, and it may choose to not disclose the truth if it does not find it prudent for those desires or beliefs to do so. At the same time, AI is incapable of promising something the same way that it means something to us.

Let us use a courtroom as an example of why AI testimony is less reliable. When a witness takes a stand, we justify our belief in what that witness says for two general reasons. The first is that it can be prosecuted for false testimony. The second is that the witness swore to tell the truth. Suppose an AI is acting as a witness as a trial; do these reasons apply to the AI? The first reason is invalid because AI cannot be prosecuted by law – at least until we make special laws for AI. But as argued in section 2.4.2.2, putting ethical blame on an AI seems absurd because it is a non-living object. The second reason is also invalid, because "swearing" something for AI means nothing, as an oath for an AI is devoid of meaning. An oath means something to us because of its semantic content and how it appeals to our morality, but for an AI, this is mere syntax. Some could say that it has syntactical value for AI, but for humans, the core value of an oath is found in its semantic content. Furthermore, the consequences of its testimony cannot appeal to its moral consciousness because it is only capable of imitation of guilt or remorse. So it seems that when an AI makes an oath, or a promise, or testimony in general, the salient part of why we believe testimony is simply not there.

Let us consider another example that highlights the difference between reasons for trust in the say-so of humans, and the say-so of AI. Suppose you go to the doctor and there are two doctors to treat you there. One of the doctors is a person, and the other doctor is an AI. Each has their own practice, and both graduated from medical school. By all accounts, they are both experts in medicine. However, they disagree about your diagnosis, and the treatment they suggest contradict each other. One of the treatments will save your life, but if you chose the wrong one – the treatment could kill you. Which one should you choose? Perhaps you ask yourself, "what makes their respective testimony reliable or trustworthy as doctors?" You reason that there are two salient reasons to trust a doctor, aside from the legal ramifications of malpractice. First of all, nobody would suffer through medical school and work the demanding hours required by doctors unless helping people meant a great deal to them. And secondly, they swore the Hippocratic oath to do no harm. Now, which doctor do you have justified reasons to believe? The answer seems to be the human doctor because the salient reasons for belief only apply to him. Indeed, the first reason, which involved the inference of the doctor's willingness to suffer for the meaning he finds in his work, does not apply to the AI because the AI cannot experience suffering or sacrifice – and more importantly, true meaning in its work. The second reason, which is inferred from the fact that the doctors swore an oath, also does not apply to the AI, because the salient content of an oath is meaningless to the AI. Thus the justifications for belief in testimony cannot be applied to the AI, and thus AI testimony cannot be deemed trustworthy for the same reasons.

It seems AI testimony cannot have the capacity for trustworthiness because it is a function of social relations. But can it be reliable? Statistically, it may seem like we are justified to expect a level of reliability, but as long as the AI is capable of deceit, we need reassurance from a human expert that can corroborate the testimony. If there are no experts who can corroborate its sayings, the testimony of an AI is nothing more than an indication that something could very well be true, but indications cannot be evidence in and of itself. I would say it is even less reliable than instrumental output, as instruments do not have the cognitive features that may complicate our relationship to its behaviour. What follows is that the epistemically engineered environment required for reliance, or trust on AI and AI*, depends on the presence of experts that can attest to

the veracity of its testimony, rather than the design dependence or societal norms that ordinarily justify a degree of trust.

2.5.3. AI testimony as an epistemic agent.

So far in this essay, I have discussed plausible justifications for belief in AI output and testimony. We have seen that many of the reasons for justification in those beliefs can be traced back to epistemically engineered environments. The question that follows from this is, what if AI testimony becomes a part of these environments? In this section, I will briefly discuss some of the problems of justifying belief in AI testimony as a part of epistemically engineered environments.

As we have seen, AI may be a rational agent and gain the epistemic status of an epistemic subject, but does that imply that we can justify belief in it as an epistemic agent? Elgin (Elgin, 2013) characterizes an epistemic agent as someone who takes their epistemic responsibilities seriously and respects the standards of epistemically engineered environments. It is someone who has the cognitive background required to support the practices, values, and methods of their field of practice. As we have discussed earlier in section 2.4.1, functionally conscious AI is unlikely to understand the meaning of the social values and attitudes such as respect that is expected from epistemological agents. Because of this, AI seems to be unqualified as an epistemic agent. Furthermore, the arguments against the reliability and trustworthiness of AI testimony suggest that we should be cautious of relying on AI testimony as a part of the division of labour that supports epistemically engineered environments. After all, an epistemic agent that is incapable of trustworthiness has the potential to damage the integrity of the environment it is implemented in.

3. Conclusion

In this essay, I have discussed how an epistemological taxonomy of AI affects plausible theories of justification for belief in its output and testimony. I have suggested a tentative epistemological taxonomy for AI as it seems that the salient epistemic features of an AI are not based on its technical properties, but rather the degree of epistemic opacity, autonomy, and complexity, and

how these features intersect. Through this discussion, I have found that for all theories of justification for belief in the reliance or trustworthiness of AI, the most important part of the justification depends on social elements. All four classes of AI depend in some way on the social norms, practices and structures that surrounds the use of AI. As AI becomes increasingly integrated into society and our daily lives, it seems increasingly important that we do not lose track of these social dimensions if we are to be justified in continued belief in AI as both an instrumental source of belief and a source of testimony.

References

Angwin, J., Mattu, S., & Kirchner, L. (2016, May 26). *Machine Bias*. Retrieved from Propublica: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Berryhill, J., Kok Heang, K., Clogher, R., & McBride, K. (2019). *Hello, World: Artificial Intelligence and.* OECD. Paris: OECD Publishing. doi:https://doi.org/10.1787/726fd39d-en.

Bostrom, N. (2017). *Superintelligence : paths, dangers, strategies.* Oxford: Oxford University Press.

Bringsjord, S., & Naveen Sundar, G. (2020, April 13). *Artificial Intelligence*, Summer 2020. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence

COMPAS Case Study: Fairness of a Machine Learning Model. (2020, September 7). Retrieved from Towards Data Science: https://towardsdatascience.com/compas-case-study-fairness-of-a-machine-learning-model-f0f804108751

Czajkowski, G. (2008, November 21). *Sorting 1PB with MapReduce*. Retrieved from Google official blog: https://googleblog.blogspot.com/2008/11/sorting-1pb-with-mapreduce.html

Defense Innovation Board. (2020). *Understanding AI Technology.* Pentagon, Department of Defense. Retrieved from https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF

Elgin, C. Z. (2013, June). Epistemic Agency. *Theory and research in education, 11*(2), 135-152. doi:10.1177/1477878513485173

Goldberg, S. C. (2017, April 29). Epistemically engineered environments. *Synthese, 197*(7), 2783-2802. doi:10.1007/s11229-017-1413-0

Herweijer, C., & Waughray, D. (2018). *Harnessing Artificial.* World Economic Forum System Initiative on Shaping the Future of Environment and Natural Resource Security / PwC / the Stanford Woods Institute for the Environment .

High-Level Expert Group on AI. (2019). *A DEFINITION OF AI: MAIN CAPABILITIES AND DISCIPLINES.* Brussels: European Comission.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI* . Brussels: European Commission.

Humphreys, P. (2009, August 1). The Philosophical Novelty of Computer Simulation Methods. *Synthese (Dordrecht), 169* (3), 615-626. doi:10.1007/s11229-008-9435-2

Jalota, R., Trivedi, P., Maheshwari, G., Ngonga Ngomo, A.-C., & Usbeck, R. (2020, January 1). An Approach for Ex-Post-Facto Analysis of Knowledge Graph-Driven Chatbots – The DBpedia Chatbot. *Chatbot Research and Design*(11970), 19-33. doi:10.1007/978-3-030-39540-7_2

Lecun, Y., Bengio, Y., & Hinton, G. (2015, May 27). Deep learning. *Nature, 521* (7553), 436-444. doi:10.1038/nature14539

Lee, P. (2016, March 25). *Learning from Tay's introduction*. Retrieved from Official Microsoft Blog: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.00000gjdpwwcfcus11t6oo6dw79gw

Matthias, A. (2004, September). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), pp. 175-183. doi:10.1007/s10676-004-3422-1

Russell, S. J., & Norvig, P. (2009). *Artificial Intelligence a Modern Approach, Third Edition* (3rd ed. ed.). Boston: Pearson.

Russell, S. J., & Norvig, P. (2009). *Artificial Intelligence, Third Edition.* Prentice Hall.

Ryan, M. (2020). In AI We Trust: Ethics, Artifcial Intelligence, and Reliability. *26*, 2749–2767. doi:https://doi.org/10.1007/s11948-020-00228-y

Savage, N. (2019, July 14). Marriage of mind. *Nature, 571*(S15), 4. doi:https://doi.org/10.1038/d41586-019-02212-4

Schneider, S. (2019). *Artificial You: AI and the Future of Your Mind* (illustrated ed.). Princeton University Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, 417-457.

Searle, J. R. (2014, October 9). What Your Computer Can't Know. *The New York Review of Books, 61*(15), p. 52.

Sparrow, R. (2007, February). Killer Robots. *Journal of applied philosophy, 24*(1), pp. 62-77. doi:10.1111/j.1468-5930.2007.00346.x

Sutrop, M. (n.d.). SHOULD WE TRUST ARTIFICIAL INTELLIGENCE? *Trames, 23*(4), 499-522. doi:https//doi.org/10.3176/tr.2019.4.07

Trust as an unquestioning attitude. (forthcoming). *Oxford Studies in Epistemology.*, 48.

Wheeler, B. (2020). Reliabilism and the Testimony of Robots. *Techné, 24*(3), 332-356. doi:https://doi.org/10.5840/techne202049123

Appendix

Figure 1.

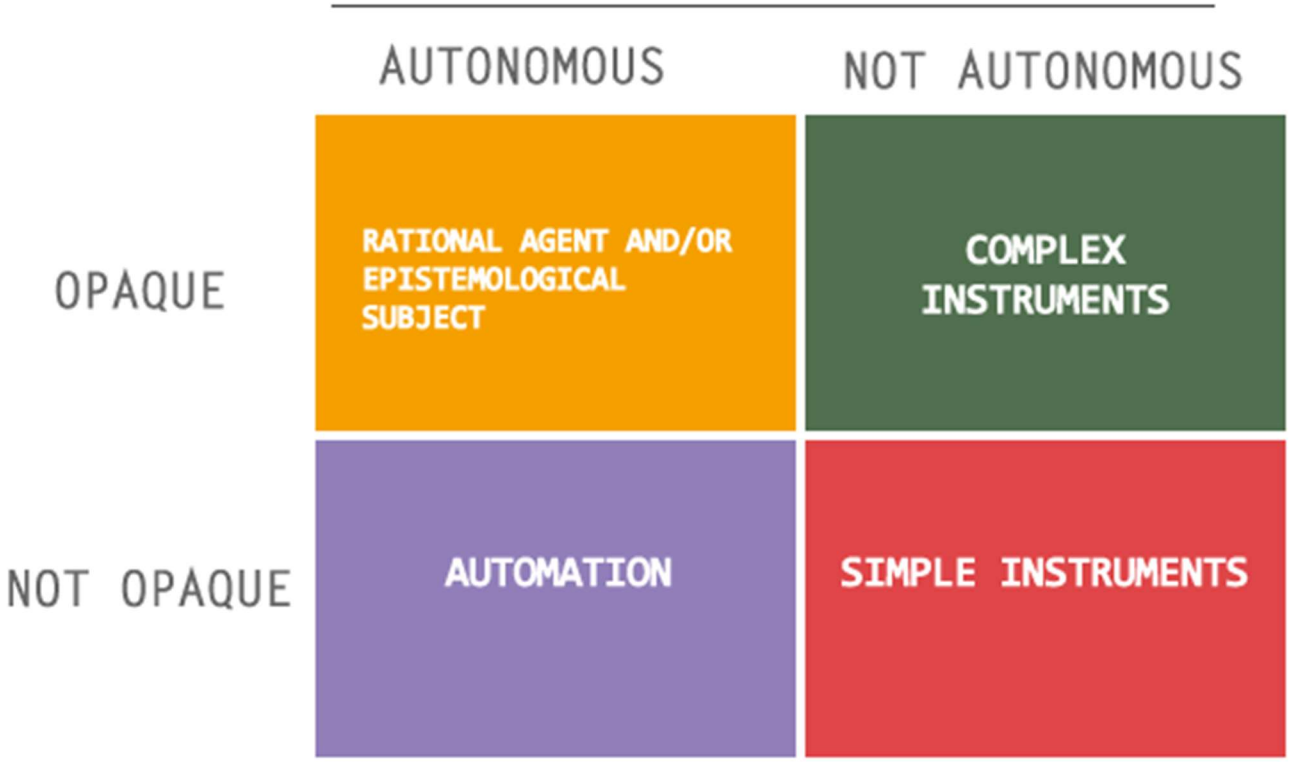

Figure 2. An impressionistic illustration of the dynamic between the epistemic features of opacity, autonomy and complexity.