Oxford Handbook of Ethics of AI: Chapter 1 summary.

I recently purchased The Oxford Handbook of Ethics of AI. This is a 900 page monster book...with small print and no pictures! (I was kind of hoping it would be a pop-up book when I saw the page count 🤣)

As a book like this is a little too much for people with a casual interest in AI and ethics, I thought I would summarise the book chapter by chapter over the next few weeks (months...).

It is important to note: I am interpreting a complex topic and trying to summarise it. I may miss key points, I may misinterpret points, so please leave a comment or reach out if you spot anything that is unclear or you think is missing!

So here is entry 1 covering the first chapter.

Chapter 1: The Artificial Intelligence of the Ethics of Artificial Intelligence. (Joanna J. Bryson)

Artificial intelligence is a field pursuing 2 related objectives: improved understanding of computer science through psychological sciences and improved understanding of psychological sciences through computer science.

Despite this Ethics was not included in either field to any great extent initially and it is only as the field expands and gains more traction that this oversight is becoming more and more apparent, and the importance of ethics in the field is being understood.

There has been a lot of effort put into thought experiments or "intuition pump", but not enough study and time has been put into what is actually computable and into the social science of how access to more information and (mechanical) computation power has altered / affected human behaviour.

So the book (The Oxford Handbook of Ethics of AI) aims to focus on the law, as this is the day to day way in which we regulate society and defend our liberties.

Intelligence is an ordinary process

Definition of intelligence: Intelligence is the ability to do the right thing at the right time. It is the ability to respond to opportunities and challenges that are presented by the current context.

Following from this, there is a section explaining how intelligence is limited by energy and or physical space. More computation requires more power and or more space, so there is a limit on intelligence based on these factors. This is why human intelligence is limited, there is only so much computational capacity that can be crammed into our head.

A further point on capacity for computation is that quantum computing may not solve this either, as although less space is required for the computers themselves, the power required for computation can be the same or even higher, depending on the type of work being done.

As a further point the author then goes on to explain that "artificial" in "artificial intelligence" is only a qualifier as to who is responsible, i.e. humans created AI, so humans are by default responsible for it.

As far as we aware, only humans can communicate about and negotiate an explicit concept of responsibility.

Over time, as we recognise consequences from actions, societies tend to assign responsibility and accountability, or credit and blame for actions, depending on whether they are viewed as positive or negative.

How we handle and deploy the use of AI and digital systems will dictate whether we obscure / diffuse responsibility or we manage to focus it in more (which in theory should be the case as we can process more information). This is one area where the law could be utilised to ensure accountability is properly apportioned (e.g. if there is a lack of such care, negligence laws could be used.)

This point needs particular thought as the definition of intelligence (and the pre-qualifier of "artificial") with debates such as "sentience, consciousness or intentionality" is often used to obscure where responsibility lies.

My key takeaways / thoughts on Intelligence is an ordinary process.
Intelligence is the ability to make the best decision possible with the current context and information.
Intelligence is limited by physical space and energy - this is why AI will be able to surpass us as it is far less constrained by these factors.

"Artificial" is only a qualifier that means "made by humans", and this makes those who make AI responsible for its actions by default.
Should we allow the negation of responsibility for AI itself as it is not conscious, or is it feasible we apportion blame to AI as it becomes significantly advanced?

AI, Including Machine Learning, Occurs By Design

AI is intentionally designed. Despite the fact that models may partially automate the process (such as training themselves on the data provided) there are still a lot of decisions such as:

- what data to train the system on
- what inputs the system has
- what outputs the system can generate
- what method is used for deciding weightings between each neuron, how each neuron learns, back propagation etc.

AI is intentionally designed and created, and they are designed to learn about regularities in data and then exploit these regularities in some way (categorisation, prediction etc.).

These decisions and processes are easy to, and should be, documented, so the designers can be held to account.

This can be thought of in a similar way to manufacturing, where the process should have clear documentation of material sourcing, processes, maintenance etc.

Essentially we can make understanding systems easier or harder to understand when building AI systems, these are all design decisions we have control over.

In the same way, the extent to which accountability and transparency are required in AI systems are also design decisions, albeit these design decisions are decided by regulators, courts and lawyers.

Given that AI is designed by people and we have a lot of control over their design, it is argued by the author that it would be sensible to focus on minimising the change to existing laws to cover AI, and instead focus on maximising compliance and relevance with existing laws.

My key takeaways / thoughts on AI, Including Machine Learning, Occurs By Design.
As AI is designed we have the ability (and possibly the responsibility) to design AI systems to comply with existing laws as closely as possible.

We should design transparency into AI systems. We should also document clearly, design decisions made, particularly in respect to training data and model validation (checking that the output is as expected).

The Performance of Designed Artifacts is Readily Explainable

Many leading AI professionals argue that if AI had to be able to explain how it worked, i.e. demonstrate "due process", this would limit innovation drastically, as most advanced machine learning techniques are far too complex to be explained.

But this argument is not correct according to the author, and they use the example of "in an audit, we do not ask for a map of an employees neurons and synapses, but instead an audit trail and witnesses that indicate appropriate processes have been followed" (paraphrased).

While fewer people may be able to "be put on the witness stand", AI allows for a standard of record keeping that potentially is far more robust and easier to achieve than relying on human recollection and documentation of process.

Any AI system can log relevant information, and maintain logs of this information. Deciding what is logged and for how long the logs are stored is up to regulators and institutions that create AI to decide.

We do not need to understand exactly how AI works, we don't need to understand torque to understand bicycle laws, we just need enough information to ensure it is used in a way that is lawful, and doing the "right things". We want to be able to demonstrate that a company complies with what it says, for example that it is not spying on it's users or putting individuals at an unfair disadvantage, or that foreign agents are unable to insert false information into a ML dataset or news feed.

We can track by who, how and when an AI system was deployed. We can also track the construction, outcomes and application of validating tests. Even a complete "black box" can be tested based on inputs and outputs and probabilities of outcomes can be calculated.

In fact, AI is far less complex to test and monitor than things like government or ecosystems, and we have decades old methods of using much simpler models to monitor more complex ones.

Finally do not assume that AI is not already regulated. All human activity, and in particular commercial activity, occurs within the context of some sort of regulatory

framework. The question is how can we optimise existing frameworks to encompass the new challenges of AI.

My key takeaways / thoughts on The Performance of Designed Artifacts is Readily Explainable.

AI and companies working in AI should be able to create some form of audit trail. Both for how it was constructed and trained, as well as for how it delivers outcomes.

The more transparent a system, the easier it is to ensure the system is acting "in the best interest" and "in alignment" with existing laws and a company's mission.

As a side thought - by designing explanations into an AI system from the start it would surely be easier to correct for biases and unexpected behaviour (e.g. "hallucinations") by knowing what clusters of neurons / nodes were activated and what the AI used to make a decision.

Intelligence Increases by Exploiting Prior Computation

Due to the limitations in instantaneous computing power, a lot of intelligence is based on past computation.

AI is advancing at a fantastic rate due in no small part to past work, which helps categorise huge datasets in a way that allows for greater learning. But with this, the good also accompanies the bad.

Stereotypes exist as they are a reflection of modern day conditions. We can all agree that some expectations that arise from this (such as a software developer being male) are not desirable and we wish to change this if possible.

But ML algorithms cannot account for this based on data alone, we have to influence this through careful curation and steering.

This leads to one theory of why AI is exploding. It isn't that we are creating new algorithms at an exponential rate, but rather that we have increased computational power and lots of new data on which to train AI.

We can expect this to plateau as there is only so much data generated (past computation) that we are able to pull from.

However, although this may plateau, the capacity for both human and AI growth will then (and has) increase(d). We have more access to information from each other (and AI will in turn have more access to information from us and other AI systems).

We all get smarter as we have access to more diverse cultural and intellectual thoughts and the minds of more people / resources.

With all that being said, the author does not believe we can be replaced with AI (for the forseable future at least), and this has a large impact on law and regulations.

My key takeaways on Intelligence Increases by Exploiting Prior Computation
The explosion in AI growth is down to increased computational power and through improved categorisation and availability of data use to train it.

Past work (past computation) has a huge impact on growth and growth potential in AI and human understanding. So at some point we will reach a plateau in this explosive growth we are seeing. Where that plateau is is an interesting thing to think about.

Stereotypes and biases are prevalent in all data, as it is a reflection on our society, how can we possibly correct for this without causing more damage?

AI Cannot Produce Fully Replicated Humans (All Models are Wrong)

Computer science is often mistaken to be a branch of mathematics, this results in people overlooking or forgetting the physical limitations of computation, e.g. time. AI is very unlikely to lead to our immortality, uploading your brain to a computer would require the computer to mirror the brain's structure, and would also involve trillions of nanosecond specific measurements.

The closest we could get is a very close abstraction, which may have its uses, but it will never be human. And building a brain from silicon will not match our phenomenological (a philosophy of experience, the human) / physical experience.

Even if we somehow managed this, technological obsolescence is measured in much shorter timescales than a typical human life. This is also true of culture. Any digital version of ourselves would quickly become out-dated in a society without constant revision.

Any digital abstraction of ourselves that is meant to contain our intelligence, as mentioned earlier, is only applicable with current context. But any abstraction we create is only a snapshot of an intelligence at a point in time, it is not the being itself.

So any abstraction (AI replica) by it's nature would be wrong almost instantaneously as it will be able to perform actions (or not perform actions, depending on it's physical capabilities) that were not feasible by the original being it was cloned / created from.

side note: that was a lot to try and process and condense, but the key point is that the second an intelligence is replicated it starts to diverge from the original due to differences in computational speed, physical abilities etc. and so will never be an accurate clone for more than a nanosecond.

Now there is "positive immortality" as a concept. For example a physical presence is not required for writing fiction, contributing to society etc. But in terms of approaching AI from the perspective of maintaining social order, from the perspective of law and regulation, it is important to draw this distinction. Any abstraction of a person then should be treated more as an extension of a person, more like intellectual property perhaps.

My key takeaways on AI Cannot Produce Fully Replicated Humans (All Models are Wrong)

Any attempt to clone ourselves into a digital brain and mechanical presence instantly means the model is outdated. Also the model will never follow the actions our biological self would follow as it will not have the same computational and physical constraints.

Because of this, any attempt to upload our consciousness should result in us treating that AI differently to that of a human, perhaps with the expectation that the human who was the model for that consciousness can consider it intellectual property, but also , possibly, also responsible for it's actions.

AI Itself Cannot be Dissuaded by law or Treaty

The primary function of the Law is not to compensate, it is to uphold social order. It does this by dissuading people from doing wrong. It makes it clear (sometimes…) what actions are considered wrong and then determines costs and penalties for committing wrongful acts.

Obviously the fines and compensation can be used to right wrongs, but this will never matter to an AI system. We can instruct an AI to not do wrong, but it will never have the aversion to jail that a being with a finite life will have.

The core problem is that laws have coevolved with humans and society, to hold humans accountable. Even corporations who are bound by laws are only fearful of them to the extent by which they affect the owners and operators of a business.

For this reason the author argues it is by necessity that a human must ultimately be held accountable for the actions of an AI. So if you, as a human, decide to let an AI operate autonomously and without oversight, it is you who is held responsible if the AI transgresses the law.

Intelligence in and of itself does not mean that the same motivations that make the law effective will work on an AI. Loneliness, social acceptance, a finite life span, social status and many other factors that make the law effective for controlling human behaviour do not apply to an AI. In fact an AI may not even find these concepts to be coherent or understandable in it's context.

My key takeaways on AI Itself Cannot be Dissuaded by law or Treaty

The fundamental problem of using the law to dissuade AI from doing wrong, is that the very factors that the law uses to control behaviour, fear of isolation, social standing, losing time from one's very short existence etc. do not matter to an AI.

So I tend to agree with the author that AI must ultimately have human oversight and a human responsible for its actions in order for laws to have any effect or even be applicable.

Otherwise if an AI transgresses the law it can simply be switched off and a new instance created almost instantaneously, meaning AI can effectively act without any fear of repercussion.

AI and ICT Impact Every Human Endeavour

In previous sections we covered that AI systems can be designed to be explainable, and that only humans can be held responsible for an AIs actions. For this reason any assertion that AI itself should be trustworthy, accountable or responsible are false.

From a legal perspective, the goal for AI is transparency in such a way that human blame can be correctly apportioned. Essentially Ai can be reliable, but it should never be trustworthy, no action taken by AI should require a "leap of faith" to determine if the action was with "good intentions" etc. Consumers and Governments should have confidence that should an AI do wrong, there is a way to determine which individual is responsible for the actions of AI used within our homes, our businesses and our governments.

AI has already helped us identify the implicit bias we may have in the language we use and that those biases are reflected in our lived realities. Due to reusing and reframing past computation, AI has highlighted and reframed information in a way that reveals a lot about our society, but it does not automatically improve us without effort.

Caliskan, Bryson and Narayan (see "semantics derived automatically from Language Corpa" https://core.ac.uk/reader/161916836?utm_source=linkout) discuss the outcome of a famous study that, given otherwise identical resumes, individuals with stereotypically African American sounding names were half as likely to be invited to interview as individuals with European American names. With this in mind, corporations are using AI to avoid implicit bias

and select diverse CVs for consideration, demonstrating that AI can, with explicit care and attention, be used to avoid perpetuating mistakes of the past.

An AI system cannot be moral as it cannot be held accountable, so we can never add our morals to a system and hope for it to act in a way that is aligned with our morals.

The problem with attempting to make an AI moral, apart from the fact that we cannot hold it accountable for its actions, is that of intention. Human intention is not easily expressed, therefore AI should be transparent and be able to receive correction from a human with oversight.

As more and more AI systems are responsible for sharing and curating information and making decisions in our society, there is also an even greater need for transparency and reporting. Otherwise those with power can grossly influence and or abuse an AI system in a way that could damage those who are vulnerable, unimportant to those in power or even that those in power consider to be a threat.

Bear in mind that “those in power” can, in the case of AI and similar systems, could simply be a coordinated effort by individuals across multiple timezones and borders, which left unchecked could also cause harm from even a subtle shift in an AIs operation.

Many people say that Machine Learning is the new AI, and that the more data we train ML on, the better / smarter the AI. Statistics teaches us that the number of data points we need to make a prediction is limited by the variation in that data, assuming a truly random sample set is produced. This is important as the only area where we actually need near unlimited / perfect data is in that of surveillance (we want to know everything then) and this should be considered when designing data sets and systems.

ICT means that we now exchange information for a reduction or elimination of cost. This “information bartering” means we surrender information to corporations in exchange for services and or products. As there is no price associated with these transactions this creates a black, or at least opaque market, reducing measurable custom and therefore tax revenue. This may be why, in some part, we are unable to see increases in productivity from AI…the transactions are not measurable in a traditional sense.

In essence: AI gives us new ways to do everything we do intentionally and a great deal more. It is not intuitive or evident the extent to which AI makes some tasks easier, and others harder. It increases and decreases the value of certain types of human knowledge and skills, social networks, personality traits and even geographic locations. It alters calculations of identity and security. It also introduces new tools to communicate and reason about all these changes and adjust for them. But it certainly makes identity, especially at the group level, more fluid and this reduces the ability to govern effectively.

My Key Takeaways for **AI and ICT Impact Every Human Endeavour**

AI can be reliable, but not trustworthy. It should be possible to extract logs and information that would allow us to apportion blame onto an individual or a corporation. This data should

be sufficient that we do not have to make a “leap of faith” on what intentions were.

Past computation has highlighted our biases. We can use AI to correct for these biases, but also it may amplify them if we are not careful.

AI cannot be moral, as to be moral an entity must be accountable, we also cannot measure “intention” in any meaningful way.

Transparency is essential to ensure that AI is not coopted or manipulated by groups, individuals or corporations for their own purposes.

We should be wary of companies and individuals who want “all the data” to improve their AI systems. Statistics shows us that there is a limit to how much data is needed to make accurate predictions (assuming a data set is truly random), and only for surveillance is there a need for “near infinite data”.

We are already in an age of “information bartering”, this means that information is shared without assigning a direct monetary value, creating a black or at least opaque market, resulting in less tax revenue. This will only accelerate with AI.

While this was a long and very dense chapter that covered a lot of points, the key takeaway for me was that of AI needing to be defensive in nature, for there to be ways to monitor it closely and ensure that “wrong doers” are not able to negatively influence its output as it has no sense of morality, and never can act morally.

Who's In Charge? AI and Governance

Humans are limited in computation by the size of their brain, so we do not change much in terms of our reaction to pain, pleasure, stress etc. Therefore our geography, our neighbours and our local environment have a large impact on us, so local governance will always be necessary to account for this.

As the world has become more global, more effort has been put into deciding what are “world level” governance issues (basic human rights for example) and which are local governance issues.

This brings us to the EU and **data governance** and extending an individual’s sovereignty to include their digital / cyberassets / personal data.

This further brings us to wealth and power. While they may not be directly related to ethics, they are heavily intertwined, as those with the wealth and power may not act in the best interests of others, in order to protect their own interests and even to try and avoid or manipulate the consensus of the law for their own gain. It is unlikely that we will solve wealth inequality and AI accountability independently.

Make no mistake, as globalisation has increased and we have experienced a reduction in the cost associated with distance (both for physical and digital assets and products), it has

allowed for an even greater wealth disparity. Local markets have become less important, competition is no longer local so there is less competition in a sector etc.

This increase in wealth disparity tends to necessitate the need for governance. One of the primary functions of government is redistribution, or at least allocation of, resources to solve communal problems. Excessive inequality can therefore be viewed as a failure of governance.

If the cost of distance is sufficiently low, then global monopolies are likely to occur.

None of these problems we are seeing today are directly because of ICT, but rather because regulatory responses were not designed for a global scale of data and service exchange.

At the same time other problems were not so much created, but rather exposed by ICT and AI (for example our global awareness of news in other Countries and regimes that may oppress their citizens, or the fact that a lot of “American (USA) problems” are now “Global problems” due to the prevalence of big tech being primarily from the USA).

One thing that has become apparent in all of this, is that humans are also heavily algorithmic. Laws can make us more so, for example with mandatory sentencing. Ordinarily though humans do have “wiggle room”.

For example, trust is cooperation with ignorance. Trust allows for cheating or innovation, sometimes it may be essential. Allowing innovation to find novel and better solutions is essential to the betterment of humanity. But the digital era means that trust, and the need for trust, are being eroded and removed. Some nations may allow the digital age to make free thought too difficult or individually risky. This will create national fragility to security threats, as well as impinging on freedom of opinion (and essential human right).

In some Countries we may seem the law move towards preserving the group, at the expense of the individual. This is not just about rights, but also robustness. Variation and individuality produce alternatives, having multiple options allows for rapid ways to change behaviour when a crisis demonstrates that a change is needed. As the digital age has fundamentally changed personal privacy, societies need to find new ways to reintroduce “wiggle room” into people’s thinking, innovation and opinion.

The author believes this would be done preferentially by allowing access to history, not destroying it, and instead acknowledging and defending individual differences, including shortcomings and the necessity of learning. But psychological and political realities remain to be explored and understood and may vary by polity (a form or process of civil government or constitution..)

My Key Takeaways for Who's In Charge? AI and Governance

I will be honest, this section seemed like a meandering mess to me, no thought felt like it led to the next, so my takeaways are more limited.

With that being said the things I took away are:

That local governance will be (and has been) eroded by globalisation, but local governance is essential due to local environmental and cultural factors.

Governance also has a key function of asset distribution and redistribution, something that will become even more important as we enter the “creation age” (my term for the AI age).

The worrying point is that I agree with the law moving towards “preservation of the group” is dangerous (and already happening). A person will end up being defined by characteristics that are not within their control, versus by their character, convictions and beliefs, and will result in greater disparity based on things like race, religion, geographic location etc. This is especially insidious with AI systems as a poorly designed AI will make decisions about an individual based on factors that may not be relevant in a specific context.

Summary and the robots themselves

To keep this article “short” this section is purely a summary of the points previously made, so I will not summarise it again here.

My takeaways

A lot to process.

I think the biggest factor that I really need to consider is that or “prior computation”. I had never framed AI or learning in that manner, yet it is an obvious (to me) point to consider as to why the pace of AI is increasing. We have more data, we have more processing power, we can keep building on an exponentially expanding growing resource of prior computation.

It also raises the question of “where is the plateau”. It happens in every industry and with every new technology. The important question to ask is “are we at the beginning of the slope or are we nearing the top” as that will certainly dictate what the future will look like and what things we need to consider for an ethics standpoint.

Additionally the need for transparency and auditability as AI is not, and cannot be “trustworthy”, only “reliable” is a key thing that I need to process and think on more.

Overall I enjoyed this first chapter, and although there were a lot of “partially formed” thoughts introduced, I feel it sets the groundwork nicely for the chapters to come!

I hope you enjoyed this summary, let me know what you think in the comments section.