Ethical Principles for Web Machine Learning

W3C Group Draft Note, 29 November 2022

▶ **More details about this document**

Abstract

This document discusses ethical issues associated with using Machine Learning and outlines considerations for web technologies that enable related use cases.

Status of this document

This section describes the status of this document at the time of its publication. A list of current W3C publications and the latest revision of this technical report can be found in the W3C technical reports index at https://www.w3.org/TR/.

This document was published by the Machine Learning Working Group as a Group Draft Note using the Note track.

This document is for guidance only and does not constitute legal or professional advice. The document will evolve and receives updates as often as needed. The whole document is open for comment and review, but input is particularly sought on Sections 3, 4 and 5.

Group Draft Notes are not endorsed by W3C nor its Members.

This is a draft document and may be updated, replaced or obsoleted by other documents at any time. It is inappropriate to cite this document as other than work in progress.

The W3C Patent Policy does not carry any licensing requirements or commitments on this document.

This document is governed by the 2 November 2021 W3C Process Document.

Table of Contents

§ 1. Introduction

> That AI will have a major impact on society is no longer in question. Current debate turns instead on how far this impact will be positive or negative, for whom, in which ways, in which places, and on what timescale. [AI4People]

> There is no 'silver bullet' here; creating technologies that will promote human flourishing and sustainable life on this planet is hard and uncertain work, involving difficult tradeoffs, some inevitable failures, and challenges that defy simple and stable solutions. But it is good work, work that can and must be done. [Vallor]

Machine Learning (ML) is a powerful technology, whose application to the web promises to bring benefits and enable compelling new user experiences. But there is also increasing awareness that ML applications can create harms, intentional and unintentional, that impact individual users, communities and society.

W3C’s mission is to “ensure the long-term growth of the web” and this is best achieved where the potential harms of new technologies like ML are considered and mitigated through a comprehensive ethical approach to the design and implementation of Web ML specifications.

As required by the charter of the Web Machine Learning Working Group, this document sets out such an ethical approach. It contains a set of ethical principles and guidance. It includes some general consideration of harms, risks and mitigations relevant to Web ML. And it offers a practical process for supplementing those general considerations with concrete risks and mitigations for specific use cases.

> NOTE: In broader debate, the terms Artificial Intelligence and Machine Learning, and their related ethical considerations (AI/ML Ethics) are often used interchangeably. Given the focus of the WG on Machine Learning, this document will generally use the terms Machine Learning or ML, and Machine Learning or ML Ethics, with the intent to refer to the broader set of issues and concerns encompassed by AI/ML.

§ 2. Machine Learning on the Web

In parallel to general advances in ML, the web platform is gaining client-side Machine Learning capabilities. Currently machine learning inference in the browser uses the WebGL graphics API, but the lack of access to platform capabilities beneficial for ML such as dedicated ML hardware accelerators constrains the scope of experiences and leads to inefficient implementations on modern hardware.

The Web Machine Learning Working Group aims to develop standards to enable access to these client-side capabilities. In web-based ML applications, the model may reside on the server or on the client, and the data processing, or inference, can be offloaded to the client.

Example use cases include person detection, facial recognition, image captioning, machine translation and noise suppression.

Web machine learning has a number of potential benefits. It could make large-scale deployment of ML systems feasible without investment in cloud-based infrastructure. This opens the door to tens of millions of do-it-yourself web developers and aligns this technology with the decentralized web architecture ideal that minimizes single points of failure and single points of control.

Local processing could also enable machine learning use cases that require low latency, such as object detection in immersive web experiences. By offloading computationally expensive tasks involving ML to on-device hardware, any web application could be enriched with ML capabilities, and existing web content progressively enhanced.

With appropriate safeguards, enabling machine learning inference in the browser (as opposed to in the cloud) *could* also enhance privacy, since input data such as locally sourced images or video streams stay within the browser's sandbox.

§ 3. General ethical issues in Machine Learning

As well as potential benefits, there is increasing awareness that the application of machine learning poses risks and can lead to harms, raising a range ethical questions. This section presents a brief overview of some key concerns.

For a general background on ethics and its relevance to ML, see Appendix 1. Background: Ethics & Machine Learning.

§ 3.1. Accuracy

The accuracy of an ML model is the proportion of examples for which it generates a correct output [Leslie]. In general high accuracy is a good thing, and low accuracy can lead to harms, for example where facial recognition systems are used in law enforcement. But highly accurate facial recognition systems can also pose risks to privacy and autonomy (e.g. mass surveillance).

In some areas such as credit-scoring or loan approval, increasing the accuracy of predictions might come at the cost of requiring access to too much personal data.

There is also concern about the over-hyping of the ability of AI to accurately predict certain things at all, particularly social outcomes such as job performance or criminal recidivism. Accuracy may be a useful measure where an area has a clear, objective ground truth (e.g. vehicle license-plate recognition) but many areas of human judgment are nuanced, messy and contextual, and simple accuracy risks being too reductive a measure.

An incorrect output produced by an ML model can have varying context-dependent impact [Leslie]. For example, when identifying cancerous skin growth, *false positives* may increase the cost of cancer detection (by requiring additional lab work to rule them out), while *false negatives* may delay treatment to the point where it is no longer effective. What matters here is lowering the risk of *false negatives*. In a judicial context, *false positives* may send innocents behind bars, while *false negatives* might make it more difficult to convict a criminal. What matters here is lowering the risk of *false positives*.

§ 3.2. Bias

Bias has a number of meanings, including a systematic deviation from a true value. This can be positive or negative. Bias is a prominent concern in ML ethics, where the concern more specifically is 'a systematic skew in decision-making that results in unfair outcomes' [CDEI].

Concerns about bias are particularly prominent where negative outcomes (such as inaccurate predictions and their consequences) disproportionately affect individuals or groups who are vulnerable or historically marginalised. Where the unfair treatment relates to protected characteristics such as race, gender, disability or sexuality, bias can constitute illegal discrimination, depending on relevant laws.

There are a number of causes of bias [Mehrabi], ranging from issues with data to algorithmic design and human perception and decision-making. Perhaps the most prominent cause is that algorithms trained to make decisions based on past data will often replicate the historic biases in that data ([Suresh] also has a useful survey of causes of bias).

§ 3.3. Fairness

There is no single definition of fairness ([Mehrabi] also has a good survey of definitions) - like ethics it is contextual and varies according to different values, perspectives and societies. But one core idea is that people should be treated equally unless there is a justified, relevant reason not to.

Fairness is often a lens through which to make sense of other ethical concerns. As noted above, bias can be positive or negative - it's when it leads to 'unfair' outcomes that it is problematic. Where ethical principles or concerns need to be balanced against each other, considering fairness often provides a guide to how to do that.

Fairness is about both outcomes and process. Outcomes should involve the fair distribution of benefits and costs, and the avoidance of unfair bias or arbitrary decisions. Procedural elements of fairness include involving communities that will be affected by ML outputs in decisions about how the systems are designed and used, and ensuring there is the ability to contest and seek redress for decisions made by ML.

Fairness could also arguably justify bias - for example biasing a system to favour people who have been historically marginalised, in order to achieve an outcome which is in some sense equal or fair (this is known as equity - treating people differently on the basis of need to achieve outcomes which are fair).

Another important aspect of fairness is the distribution of access to computationally complex ML approaches, and the benefits that come from access. People living in countries with less powerful or functioning infrastructure, or who cannot access sufficient computing power, may be unfairly disadvantaged.

§ 3.4. Safety & Security

Safety includes that an ML system should be accurate, but also that it should be reliable (perform as intended, and continue to do so over time), secure (against adversarial attacks), and robust enough to do these things in real-world, unpredictable and sometimes challenging conditions ([Leslie])

Safety is a broad concern, but is particularly relevant where the failure of ML systems could result in real-world harm - for example with medical diagnosis or self-driving cars.

There are a number of security risks to machine learning, including training data poisoning, adversarial inputs, or model inversion and adversarial inference attacks which can expose model parameters or training data ([Xue]).

Machine learning can also increase the effectiveness of other types of security attacks, for example by enabling more effective impersonation for social engineering and phishing attacks.

§ 3.5. Privacy

There are a number of ways in which ML systems can pose risks to privacy.

One is where systems that undermine privacy operate without a user's knowledge or explicit, informed consent. This is true of systems that undermine privacy explicitly (surveillance systems), but also where undermining privacy is a potential byproduct of intended, legitimate use (e.g. if an ML system which has access to a user's video camera).

There are also privacy concerns about the data used to train models. Data may be collected in a way which violates privacy, such as without consent from users (e.g. scraping personal information). Models may 'leak' personal data (e.g. large language models [Weidinger]). Legitimately collected data may also be compromised, for example through reverse engineering or inference style attacks which can de-anonymise model training data.

The accuracy of the predictions of ML systems may also present risks. Just as the outputs of sensor APIs could be used to identify, fingerprint or correlate user activity (e.g. if the output is too precise), it is possible that the outputs of ML systems could pose similar risks.

And use of ML systems to infer sensitive, personal data about users based on non-sensitive data (e.g.

inferring sexuality from content preferences) may also violate privacy.

Some jurisdictions (e.g. EU/GDPR) also provide a ‘right to be forgotten’, which arguably could include being removed from ML training data. So a privacy-protecting approach would need to ensure that appropriate processes and technical capabilities are in place for this to happen (see e.g. [Bourtoule]).

§ 3.6. Transparency

Very broadly, transparency is about users and stakeholders having access to the information they need to make informed decisions about ML. It’s a holistic concept, covering both ML models themselves and the process or pipeline by which they go from inception to use. [Vaughan] (following the [EGTAI]) propose 3 key components:

- **Traceability**: Those who develop or deploy machine learning systems should clearly document their goals, definitions, design choices, and assumptions.
- **Communication**: Those who develop or deploy machine learning systems should be open about the ways they use machine learning technology and about its limitations.
- **Intelligibility**: Stakeholders of machine learning systems should be able to understand and monitor the behavior of those systems to the extent necessary to achieve their goals.

Understanding ML systems involves two key related concepts [Gall]:

- **Interpretability**: is about the extent to which a cause and effect can be observed within a system.
- **Explainability**: the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms.

Lack of interpretability and explainability is known as the black-box problem, which is particularly prevalent with more complex ML approaches such as neural networks.

§ 3.7. Accountability

Given that ML systems are increasingly being used in high impact areas (healthcare, welfare, criminal justice) and that harms can be large when they go wrong, and that actors in the ML pipeline take responsibility for considering the impact of ML systems, and accountability for when things go wrong.

“Algorithms and the data that drive them are designed and created by people – there is always a human ultimately responsible for decisions made or informed by an algorithm. "The algorithm did it" is not an acceptable excuse if algorithmic systems make mistakes or have undesired consequences, including from machine-learning processes.” [FATML]

Transparency is an enabler for accountability (we need to be able to see what is going wrong and where to be able to determine responsibility). It also requires proper processes for the consideration of risks to be in place, documentation of policies and processes, and the means for those who are harmed to seek redress. The developers of ML systems should also take responsibility for any 3rd party ML they use in their system.

Increasingly in some jurisdictions, there are formal legal mechanisms for accountability and seeking redress.

§ 3.8. Human Control and Decision-making

The need for accountability, as well as other concerns above such as accuracy and fairness, have led to the assertion of the importance of humans making in the final decision in high stakes applications. More broadly, ML applications should always be under ultimate human control.

But there are pitfalls too where ML approaches support human decision-making - problems with explainability can inhibit the full exercise of human capabilities, or humans may exhibit “automation bias” where they place too much trust in information or recommendations provided by an ML system.

§ 3.9. Environmental Impact & Sustainability

There is increasing awareness that computationally complex ML approaches trained on very large data sets can have a large environmental impact, given the amount of energy required to power the training phase.

The broader concern with sustainability suggests that ML applications and systems should not undermine the sustainability of the physical, social and political ecosystems in which they’re deployed. This might include the impact on jobs, employment and the economy, or on the quality of and access to information necessary for a functioning democratic system.

§ 3.10. Types of harm

The above list contains some potential sources or causes of harm from machine learning. It is also important to be aware that harm can take a number of different forms, all of which should be considered.

As noted above, harms can impact individuals, groups and society. To take the example of a biased facial recognition system [Smuha]:

- this may lead to wrongful discrimination against an **individual** (e.g. wrongful arrest).
- where a number of individuals who belong to a **group or collective** suffer this discimination (e.g. because of shared ethnicity), there is a group harm. This could be the sum of the individual harms, as well as harms such as an increase in prejudice towards that group caused by the perpetuation of historic bias.
- here could be a harm to the interests of **society**, such as being able to 'live in a society that does not discriminate against people based on their skin colour and that treats its citizens equally.' [Smuha]

Harms can also take a number of forms. These can include:

- **physical**, either directly (e.g. the failure of driver-less cars), or indirectly (e.g. flaws in a system leading to incorrect medical diagnosis).
- **allocative**, when a system unfairly allocates or withholds from certain individuals or groups an opportunity or a resource (e.g. benefits or loans) [Crawford].
- **representational,** when systems "reinforce the subordination of some groups along the lines of identity." [Crawford] e.g. when a Google search for 'CEO' returns mostly pictures of white men, or image recognition systems generate offensive labels for people of colour.

§ 4. Ethical Principles for Web ML

The following ethical values and principles are taken from the UNESCO Recommendation on the Ethics of Artificial Intelligence [UNESCO]. They were developed through a global, multi-stakeholder process, and have been ratified by 193 countries. There are four high level values, and ten more detailed principles, to which we've added an additional, explicit principle of 'Autonomy'. For more on why these have been adopted, see Appendix 2. Why the UNESCO principles were chosen.

These values and principles should drive the development, implementation and adoption of specifications for Web Machine Learning. They include guidance (adapted from UNESCO and W3C sources) which provides further detail on how the values and principles should be interpreted in the W3C web machine learning context.

The following terms are used:

- 'ML actors' refers to stakeholders involved in web ML: specification writers, implementers and web developers

- ‘ML systems’ refers to the ML model or application that is making use of web ML capabilities

The next section (S.5) provides further guidance on how to operationalize the principles and turn them into specific risks and mitigations.

§ 4.1. UNESCO Values

These indicate desirable behavior and represent the foundation of the principles

§ VALUE 1) Respect, protection and promotion of human rights and fundamental freedoms and human dignity

ML actors and systems should treat all human beings as being of equal worth, and no individual, group or society should be harmed. ML systems should be designed in a human-centric way, to promote human flourishing, and should respect and enhance human autonomy. This includes enabling meaningful agency, control and choice.

ML actors should not enable state censorship, surveillance or other practices that seek to limit these freedoms. Nor should they enable manipulation, misinformation, harassment or persecution.

§ VALUE 2) Environment and ecosystem flourishing

Environmental and ecosystem flourishing should be recognized, protected and promoted through the life cycle of ML systems. ML actors should reduce the environmental impact of ML systems, including but not limited to their carbon footprint, to ensure the minimization of climate change and environmental risk factors, and prevent the unsustainable exploitation, use and transformation of natural resources contributing to the deterioration of the environment and the degradation of ecosystems.

§ VALUE 3) Ensuring diversity and inclusiveness

Respect, protection and promotion of diversity and inclusiveness should be ensured throughout the life cycle of ML systems. This may be done by enabling the active participation of all individuals or groups, regardless of identity, lifestyle choices, beliefs, opinions, expressions or personal experiences. This should include the meaningful option not to use ML systems. ML users should not be disadvantaged because they lack necessary technological infrastructure, education or skills.

§ VALUE 4) Living in peaceful, just and interconnected societies

ML systems should not segregate, objectify or undermine freedom and autonomous decision-making or the safety of human beings and communities. They should not divide and turn individuals and groups against each other, or threaten coexistence between humans, other living beings and the natural environment. ML systems should be built to cross regional and national boundaries.

§ 4.2. UNESCO Principles

These unpack the values underlying them more concretely so that the values can be more easily operationalized.

§ PRINCIPLE 1) Proportionality and Do No Harm

When developing ML systems, developers should consider what harm the application could do to individuals, groups and society, especially to vulnerable people. They should seek to eliminate or minimise those harms.

The use of ML and any data gathered to enable it must be proportional to achieve a given legitimate aim. It should not violate human rights, or be used for social scoring or mass surveillance purposes.

ML actors should prioritize potential benefits for users over potential benefits to developers, content providers, user agents, advertisers or others in the ecosystem, in line with the priority of constituencies.

§ PRINCIPLE 2) Fairness and non-discrimination

The benefits of ML systems should be available and accessible to all.

ML actors should minimize and avoid reinforcing or perpetuating bias and discrimination, particularly against vulnerable and historically marginalised groups. This includes but is not limited to bias based on gender, gender identity and gender expression, sexual orientation, disability (both visible and invisible), mental health, neurotype, physical appearance, body, age, race, socio-economic status, ethnicity, caste, nationality, language, or religion.

ML actors should ensure that the outcomes of ML applications are fair, and that effective remedy is available against discrimination and biased algorithmic determination. ML systems should work equally well for all users, and where this is not true for 'outliers', fallback solutions should be provided.

This principle also applies to access to web ML systems: they should be fully accessible to people with disabilities, appropriately localized, and accommodate people on low bandwidth networks and with low specification equipment.

For guidance on digital accessibility, see the Introduction to Web Accessibility published by the W3C Education and Outreach Working Group.

§ PRINCIPLE 3) Autonomy

ML systems should respect and enhance human autonomy. This includes enabling meaningful agency, control and choice. Users should give informed consent before Web ML is used.

ML systems could be used to manipulate and deceive people, complicate isolation, and encourage addictive behaviors. ML actors should mitigate against these potential abuses and patterns when creating ML systems, and avoid introducing technologies that increase the chance of people being harmed in this way.

§ PRINCIPLE 4) Right to Privacy, and Data Protection

ML systems should be designed and implemented to ensure that privacy and personal information is protected throughout the life cycle of the application. This includes training data - ML models should not be used if their training data has violated privacy. ML actors should ensure they are accountable for this, and should conduct adequate privacy impact assessments, and implement privacy by design approaches.

Data should be collected, used, shared, archived and deleted in ways that are consistent with local and international law.

For further guidance on privacy, see [security-privacy-questionnaire].

§ PRINCIPLE 5) Safety and security

ML systems should actively support safety and security. Unwanted harms (safety risks), as well as vulnerabilities to attack (security risks) should be avoided and should be addressed, prevented and eliminated throughout the life cycle of ML systems to ensure human, environmental and ecosystem safety and security. ML actors should ensure that systems are accurate, reliable and robust over their entire life-cycle, and should make sure that users understand any risks they are taking when using an application.

See also [security-privacy-questionnaire].

§ PRINCIPLE 6) Transparency and explainability

ML actors and systems should support transparency and explainability. It should always be possible to determine how a web ML application was built and how the code works, in line with the "view source" ethos of the web.

Users and third-parties such as civil society groups and researchers should be able to audit and inspect ML systems for security, privacy, bias, fairness and other ethical concerns outlined in these principles.

ML actors should inform users when a product or service is provided directly or with the assistance of ML, and users should be fully informed when a significant decision is informed by or is made on the basis of ML. They should be able to access the reasons for a decision affecting their rights and freedoms, and ML actors should provide means for users to request review and correction of these decisions.

ML systems and outcomes should be explainable. ML actors should also provide clear explantions of the benefits and risks of different ways of accessing ML capabilities (e.g. client vs server-side inference). ML actors should promote tools and approaches that enhance explainability and meaningful user control.

ML actors should also be transparent about the steps they have taken to consider and implement these ethical principles.

§ PRINCIPLE 7) Responsibility and accountability

Appropriate oversight, impact assessment, audit and due diligence mechanisms should be developed to ensure accountability for ML systems and their impact throughout their life cycle.

It should always be possible to attribute ethical responsibility and liability for the outcomes of ML systems, and decisions and actions based on them, to people or organisations corresponding to their role in the life cycle of the ML system.

ML actors should take responsibility for third-party ML models and approaches used in their systems.

§ PRINCIPLE 8) Sustainability

ML actors should ensure that the human, social, cultural, economic and environmental impact of ML applications is sustainable.

They should develop and favour approaches which minimise power consumption, data storage and processing requirements, and maxmimise the lifespan of physical devices through maintaining compatibility.

§ PRINCIPLE 9) Human oversight and determination

In scenarios where decisions are understood to have a high impact or one that is irreversible or difficult to reverse, or may involve life and death decisions, final human determination should apply.

The decision by users to cede control and decision-making power to an ML system should be explicit, fully informed and limited to a specific context.

§ PRINCIPLE 10) Awareness and literacy

ML actors should ensure that all members of society have access to the information they need to make informed decisions about their use of ML systems.

§ PRINCIPLE 11) Multi-stakeholder and adaptive governance and collaboration

The participation of different stakeholders throughout the life cycle of ML projects is necessary to enable the benefits to be shared by all, and to mitigate risks.

Where tensions arise between the principles above, they should be resolved through democratic stakeholder engagement and participation.

Web ML specifications and systems should support open standards and interoperability.

§ 5. Operationalization: Putting the Principles into Practice

The principles outlined above help map out the major areas of ethical concern, and the guidance starts to fill in some detail. But by themselves, the principles and guidance risk being too abstract and achieving nothing concrete.

What matters most for making any approach to ML ethical is to operationalize the principles - to turn them into concrete actions. This section offers practical guidance on how to do that.

One way to make an ethical approach concrete is through process. This often takes the form of checklists to consider and apply throughout the life-cycle of developing an ML system, and covers aspects such as ensuring diversity on teams and consultation with stakeholders as well as areas more specific to the technical development of the ML system.

In time this document may evolve to become a checklist for ML actors to work through. For now, those looking for an example of this sort of process guidance could find numerous examples on the web - one worth looking at is the ICO AI and data protection risk toolkit. The ICO is the UK's information rights and data privacy regulator. Their toolkit aims to help actors comply with UK Data protection law (similar to GDPR) when building AI. It contains a useful checklist of practical steps to mitigate risk.

As well as the full-cycle process, perhaps the most important single activity in developing ML systems ethically is thinking through the specific risks and mitigations for any proposed approach. There are a number of different ways to do this, including Consequence Scanning, Harms Modelling and Algorithmic Impact Assessments (e.g. this Candian government AIA).

At its simplest, the core approach is to take known areas of concern - as laid out by the principles and guidance - and for each of them think through what could go wrong (consequences/risks/harms), and what measures could be put in place to prevent that or minimise impact (mitigations).

So for example, if we consider the principle of "Fairness and non-dicrimination", a risk might be that biases in training data lead to model predictions that are less accurate for particular groups, resulting in real-world harms (e.g. denial of services). A mitigation might be to test the training data and model properly for bias, or indeed it might be to not make access to essential services dependent on ML systems.

> Note: To provide a more structured approach to operationalization, we have developed a workshop format that can be adapted and used to help identify such risks & mitigations.

It's worth noting that the best tools for operationalization might depend on an actor's role in the ML eco-system - for standards writers, specific thinking about risks and mitigations will be most useful. For developers, responsible ML checklists covering the product life-cycle will also be important.

§ 5.1. What about when there's conflict between principles or the interests of stakeholders?

In ethics there are often no easy answers. Ethical problems often don't have neat, permanent solutions. Principles may be in conflict or tension with each other - for example having more data about non-typical users to tailor solutions to them (for fairness purposes) might come at the expense of privacy. The views of different stakeholders affected by any ML system will need to be balanced. And those views and wider values will change and evolve over time.

On balancing all these things, UNESCO says:

> In any given situation, a contextual assessment will be necessary to manage potential tensions, taking into account the principle of proportionality and in compliance with human rights and fundamental freedoms … To navigate such scenarios judiciously will typically require engagement with a broad range of appropriate stakeholders, making use of social dialogue, as well as ethical deliberation, due diligence and impact assessment.

As this suggests, referring to the higher level values such as fundamental human rights and freedoms might help. Fairness is often a useful guide to balancing competing interests and values. Bear in mind W3C's priority of constituencies.

And process is often as important as outcome. The principles in this document don't just cover what issues you should think about (such as bias) but also importantly suggest how an ethical process should be run - ideally it should involve diverse participants, consult affected stakeholders, be open, democratic and transparent (so keep a record of the process and ideally make the results publicly available) and open to contestation.

§ 6. Register of Risks and Mitigations

> Note: The risk register is work in progress and welcomes further review and tidying up.

ML actors are encouraged to go through their own process of thinking through the risks and mitigations for any ML system they are developing. The wide range of possible use-cases, and the rapid pace of development of ML technology, mean that any pre-existing list of risks and migitations will never be complete. Mitigations will also vary according to an ML actor's position in the ecosystem - developers will have different responsibilities and influence than specification writers.

However, such a list is useful for a number of reasons. It can save time and re-inventing the wheel, and also allow for best practice to be captured and shared.

So this section is for gathering risks and mitigations as they are identified, and in time should develop into a register of key Web-ML risks and mitigations.

§ 6.1. Proportionality and Do No Harm

§ 6.1.1. Risks

Risk PDNH-R1

Malicious apps are easier to accidentally launch on the Web (trying to think about how ML on the Web is different from ML in Android/iOS apps or installed Windows/MacOS apps or via remote API calls.)

Risk PDNH-R2

Might be used by malicious actors to hijack CPU.

Risk PDNH-R3

Might be used for more sophisticated manipulation of people and their attention.

§ 6.1.2. Possible Mitigations

Risk PDNH-R1 Mitigations

Note: Contributions welcome via GitHub.

Risk PDNH-R2 Mitigations

Note: Contributions welcome via GitHub.

Risk PDNH-R3 Mitigations

Note: Contributions welcome via GitHub.

§ 6.2. Fairness and non-discrimination

§

6.2.1. Risks

Risk FND-R1

Scaling up ML via browsers creates risks of scaling up bias issues linked to ML training.

Risk FND-R2

ML approaches optimize for the majority, leaving minorities and underrepresented groups at risk of harm or sub-optimal service (see e.g. Treviranus).

Risk FND-R3

Differences in Internet connection speeds across geographical locations and large size of production-grade models means the user experience of on-device inference is not equal in all locations.

Risk FND-R4

Speech recognition must recognize different accents, including regional, ethnic, and "accents" arising from a person's disability - a focus on "mostly fair but left out the edges" will result in massive discrimination.

Risk FND-R5

Bias in ML training can a) make ML non-useful to some people by effectively not recognizing their personhood, or b) interfere with ability to conduct tasks efficiently, effectively, or at all, or c) create a new digital divide of ML haves and have-nots.

Risk FND-R6

That the WebML Working Group has very little control over models … is it able to influence those who do build them enough to ensure this principle is operationalised.

Risk FND-R7

Imagine doing ML-based captions: this raises issues about accuracy, efficiency, but also burden-shifting: if the captioning is happening on the local device, it may create burdens for the people that are the least able to change it while being the typical target.

Risk FND-R8

One cannot rely on simple classifications of individuals into homogeneous social groups (e.g., binary gender ca categorizations that exclude non-binary individuals). In particular, disability is characterized by diversity, and not by any property that distinguishes people who have from those who do not have disabilities.

Risk FND-R9

There are also important issues of "proxy discrimination" that have been brought out in the literature, and which should be considered (i.e., machine learning systems that discover protected classes of persons even in cases in which such classifications and obvious proxies for them are excluded from the data used in training).

Risk FND-R10

One example is how geographic pricing is employed by companies like Amazon – e.g. depending on where your IP address is located to or depending on your device type, different prices are presented to shoppers – it's a question if this is unethical or unlawful, but it is something that is happening and also begs to question whether or not this is something we want to speak to → taken further you can also imagine reducing service based on geography or hardware or other factors in a way that is automated through ML systems.

§ 6.2.2. Possible Mitigations

Risk FND-R1 Mitigations

Browser-assisted mechanisms to find out about the limitations and performance characteristics of ML models used in a Web app. This could build on an approach published in Model Cards for Model Reporting where making this report machine-discoverable would allow for the web browser to offer a more integrated user experience. Another transparency tool is the Open Ethics Transparency Protocol.

Risk FND-R2 Mitigations

ML actors should provide fallback solutions for these inevitabilities.

Risk FND-R3 Mitigations

This issue is not specific to ML and can be mitigated in part by using a Content Delivery Network and by offering reduced size models.

Risk FND-R4 Mitigations

> Note: Contributions welcome via GitHub.

Risk FND-R5 Mitigations

> Note: Contributions welcome via GitHub.

Risk FND-R6 Mitigations

> Note: Contributions welcome via GitHub.

Risk FND-R7 Mitigations

> Note: Contributions welcome via GitHub.

Risk FND-R8 Mitigations

> Note: Contributions welcome via GitHub.

Risk FND-R9 Mitigations

> Note: Contributions welcome via GitHub.

Risk FND-R10 Mitigations

> Note: Contributions welcome via GitHub.

6.3. Autonomy

6.3.1. Risks

Risk A-R1

> ISSUE 1 Identify risk mapping for corresponding mitigation.

Risk A-R2

That browsers will cease to be *user agents*. Autonomy is a key differentiator for the web vs. alternative content and app platforms.

Risk A-R3

Users have lesser and lesser control on what we see and who sees us. We're tracked by 1st and 3rd parties and we see what others want us to see (e.g. ads). Hence, based on the the principle: people should be able to render content as they want, not only should ML systems take care of that, but also help in countering this global problem.

Risk A-R4

Black boxes of ML models might negatively impact the ability of Web Extensions to bring more control (and thus autonomy) to end-users for their experience on the Web.

Risk A-R5

Web accessibility can enhance individual autonomy, by making more aspects of life "self-serve". It can also destroy autonomy, by designing only for the middle and "leaving others out in the cold" as society adopts the ML over other ways of accomplishing objectives.

Risk A-R6

ML in the Web could very well be used to enhance user's capabilities by acting as an assistant in a privacy preserving way. I.e. generating calendar events from emails or websites without sending information back up to the servers.

It could erase their autonomy by being used against them by the website using the ML as a gatekeeper before providing human access. I.e. How chatbots are used today.

Users would be wary to give consent when something which can help them can be equally used to control them or restrict access.

Risk A-R7

Cannot fully enforce informed consent requirement for the web for ML. E.g. inference with generic WebGL/Wasm capabilities possible without consent, even if purpose-built APIs would require informed consent.

Risk A-R8

Web ML systems are used without informed user consent.

Risk A-R9

Browser standards like MV3 makes the implementation harder.

Risk A-R10

Example: ML / IOT devices will be used with the intention of increasing autonomy of e.g., aging people, people with disabilities, etc., but have the risk of instead reducing autonomy if it’s not usable as designed to some users due to bias etc.

Risk A-R11

Corporate priorities will constantly be against user choice (autonomy), things like making it very difficult to choose a different option than the corporation wants users to make could easily become worse in ML scenarios.

Risk A-R12

Function creep - that a user consents to data / access / use of ML in one context, but then the use is extended beyond that context without explicit consent.

Risk A-R13

Permission / Decision fatigue is another risk, if we ask people to explicitly allow every new web feature that could be abused. It’s a hard tradeoff. By asking for explicit decisions, we might actually reduce the chance that people are making informed decisions, because they are cognitively overloaded and don’t have the time or mental energy to really understand the implications.

Browsers today do not ask user consent for things like JavaScript usage, Wasm, WebGL, WebGPU, web workers, etc. All of those can be used to perform “ML”.

Risk A-R14

Permission/consent should definitely be sought from users when accessing sensitive information about their computer or environment. Cameras, Bluetooth devices, microphones, location, controllers, gamepads, and XR devices should all be under permission prompts.

Risk A-R15

Does this include informing users about the capabilities and limitations of the system, as well as the associated risks? Informed choice needs to be guided by an understanding of capabilities, limitations, and how the system should fit into the social context in which it is intended to be used.

Risk A-R16

People might feel that their trust is betrayed if they don’t know what a web app is doing with their data. This isn’t specific to Web ML, perhaps, but it’s more salient, or more in the news.

It can be hard to explain why someone might want to enable Web ML. Eg, it’s actually safer, because your personal data will remain on your device and won’t be sent to remote servers. You’ll have a better experience or new features in the web app.

§ 6.3.2. Possible Mitigations

Risk A-R1 Mitigations

Similarly to videos, the sites should make it opt-in to load large models on load or run expensive compute tasks.

Risk A-R2 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R3 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R4 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R5 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R6 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R7 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R8 Mitigations

Things that *end users* could be asked to do…

- Growing awareness about the risks (like phishing)
- Surfacing the value of data that users share for ML can be re-used in different contexts (e.g. legal, commercial)

Things that *developers* could be asked to do…

- Develop guidance for ethical ML that includes bringing user awareness
- Open source ML algo - auditability / certification

Things that *implementers* could be asked to do…

- Upstream frequent ML-built features into browser features where they can be used in a clearer/less UX intrusive framework (as an incentive towards the safest approach)
- For a purpose-built APIs, the browser could make the usage detectable (e.g. via a web extension)
- Linked to incentive
- If ML has been certified as quality and privacy good. Or rated (A-F?) users could choose to only enable ML features at a certain level.

Things that *regulators* could be asked to do…

- Quality Assurance certificates for the algos,

Things that *standard makers* could be asked to do…

- Develop best practice guidelines for devs

Things that *no one* can fix or control…

- Developers giving users trivial incentives to load a data leaking ML model. Silly hats for all of your data.

Risk A-R9 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R10 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R11 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R12 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R13 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R14 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R15 Mitigations

> Note: Contributions welcome via GitHub.

Risk A-R16 Mitigations

> Note: Contributions welcome via GitHub.

§ 6.4. Right to Privacy, and Data Protection

§ 6.4.1. Risks

Risk RPDP-R1

> ISSUE 2 Identify risk mapping for corresponding mitigation. ¶

Risk RPDP-R2

Fingerprinting of various kinds: disability

Risk RPDP-R3

One area we came across recently related to ML is in the context of the WebXR raw camera access API. The API could allow raw access to the camera image (vs the regular AR API that only exposes room geometry). This allows for more functionality but puts the user at risk - for example camera image could be piped to a ML subsystem which is doing facial recognition outside of user's consent. Documented in our TAG issue.

The wider issue is that ML as a 1st class feature on the Web creates additional risks for existing APIs (such as camera access).

Risk RPDP-R4

Addition of ML creates additional risks for use of existing APIs that were not present previously.

Risk RPDP-R5

Different jurisdictions have different regulations for data protection and rights to privacy. Demonstrating that your model is consistent with one or another could be confusing.

Risk RPDP-R6

ML models may be based on training data that abused privacy.

Risk RPDP-R7

It will be necessary to obtain data from marginalized individuals (including those who are unable to give informed consent themselves) in order to ensure they are not discriminated against and that they are included in the product, but their data need to be treated carefully and respectfully, and there are issues of consent involved. Under what circumstances can others give consent on their behalf? Consider, for example, people with certain cognitive disabilities who cannot give voluntary, informed consent to a particular data collection activity.

Risk RPDP-R8

Sites could claim compliance with relevant laws and principles without actually being compliant. (Transparency and third-party auditing are important here.)

Risk RPDP-R9

Fingerprinting systems uses hardware accelerated ML APIs to improve their tracking capabilities.

Risk RPDP-R10

Doing processing on user’s device could be good for privacy, but could also be excuse to shift cost of computation to the end user.

Risk RPDP-R11

Another risk is that people distrust and turn off Web ML, and the alternative is worse, from a privacy perspective. The web app can still use ML, but may do so by sending private data to remote servers that are less secure than the local device.

§ 6.4.2. Possible Mitigations

Risk RPDP-R1 Mitigations

Requiring explicit consent to access privacy-sensitive capabilities such as on-device camera.

Risk RPDP-R2 Mitigations

> Note: Contributions welcome via GitHub.

Risk RPDP-R3 Mitigations

> Note: Contributions welcome via GitHub.

Risk RPDP-R4 Mitigations

> Note: Contributions welcome via GitHub.

Risk RPDP-R5 Mitigations

> Note: Contributions welcome via GitHub.

Risk RPDP-R6 Mitigations

> Note: Contributions welcome via GitHub.

Risk RPDP-R7 Mitigations

> Note: Contributions welcome via GitHub.

Risk RPDP-R8 Mitigations

> Note: Contributions welcome via GitHub.

Risk RPDP-R9 Mitigations

> Note: Contributions welcome via GitHub.

Risk RPDP-R10 Mitigations

> Note: Contributions welcome via GitHub.

Risk RPDP-R11 Mitigations

> Note: Contributions welcome via GitHub.

§ 6.5. Safety and security

§ 6.5.1. Risks

Risk SS-R1

Possible to leak the locally stored data, even sensitive data such as biosignature？

What kind of capabilities would the ML system get and thus leak sensitive local data?

Risk SS-R2

Model drift - that a model stops performing as well as real world data diverges from training data over time.

Risk SS-R3

Censorship requirements of governments and other actors, if operationalized into ML, create massive risks for individuals as well as societal evolution - ranging from being unable to accomplish objectives that our principles say they should, to “being tattled on” to the autocrats and suffering real-world retaliation.

Risk SS-R4

A model can produce results that are blindly trusted. If the model is open to compromise, it will produce inaccurate results, which can be influential.

An example could be of an app that is intended to “help you cross the street” as a visually limited person - but if that application fails to detect a cyclist or car, then you could create physical harm to the user of that application.

Risk SS-R5

ISSUE 3 Identify risk mapping for corresponding mitigation. ¶

§ 6.5.2. Possible Mitigations

Risk SS-R1 Mitigations

Note: Contributions welcome via GitHub.

Risk SS-R2 Mitigations

Note: Contributions welcome via GitHub.

Risk SS-R3 Mitigations

Note: Contributions welcome via GitHub.

Risk SS-R4 Mitigations

Note: Contributions welcome via GitHub.

Risk SS-R5 Mitigations

Can be at least partially mitigated by transparency and third-party auditing.

§ 6.6. Transparency and explainability

§ 6.6.1. Risks

Risk TE-R1

Web developers are familiar with developer tools integrated into browsers used to inspect HTML, CSS and JavaScript. These developer tools, however, do not currently understand neural network models and model inspection requires specialized tools. This complicates development and raises the barriers to entry for web developers venturing into machine learning.

Risk TE-R2

Complexity is the enemy of transparency. ML models are complex and getting more complex over time.

Risk TE-R3

ML “closed boxes” doing something out side of users’ control and understanding and the browser not able to audit or control or otherwise warn the user.

Risk TE-R4

Transparency may be operationalised in a way which doesn’t make sense to users and doesn’t respect autonomy and allow them to make informed decisions.

Risk TE-R5

The difficulty of explaining Web ML’s benefits and drawbacks may lead people to make choices that are worse for them. Eg, they might turn off Web ML, not understanding that it’s better for privacy to keep the data local. (I’m thinking here about the transparency and explainability of the API, not the ML model.)

§ 6.6.2. Possible Mitigations

Risk TE-R1 Mitigations

Web APIs by their design make it possible to integrate into browsers developer tools features that help build intuition on how neural networks work, in the spirit of "view source" principle.

Risk TE-R2 Mitigations

Web-based visualization tools have been developed for deep networks for educational use and their integration into browsers remains further work.

Integrate into web browser developer tools a conceptual graph of the model's structure to inspect and understand the model architecture.

The ML model could be viewed by an integrated tool and in a visual way such Netron does today.

Risk TE-R3 Mitigations

> Note: Contributions welcome via GitHub.

Risk TE-R4 Mitigations

> Note: Contributions welcome via GitHub.

Risk TE-R5 Mitigations

> Note: Contributions welcome via GitHub.

§ 6.7. Responsibility and accountability

§ 6.7.1. Risks

Risk RA-R1

During the discussion around DRM on the Web via Encrypted Media Extensions, a lot of focus was on whether security researchers would get protected in case they reverse-engineered DRM systems on the Web (which was seen as a net good for the Web, but a legal risk for researchers); a similar challenge may arise for ML models as they get reviewed against e.g. bias.

Risk RA-R2

Assuming long-tail web developers will prefer to use 3rd party ML models due to cost of training your own (similarly to JS frameworks in general). This means the ethical responsibility and liability is deferred (in part?) to the 3rd party.

Risk RA-R3

The use of 3rd party models introduces an external dependency to a possibly critical component of the web (app) experience.

Risk RA-R4

We can’t force developers to follow these principles and guidelines.

Risk RA-R5

That the WebML Working Group has very little control over models … is it able to influence those who do build them enough to ensure these principles are operationalised.

Risk RA-R6

ISSUE 4 Identify risk mapping for corresponding mitigation. ¶

Risk RA-R7

ML models can operate as black boxes, and when integrated in a platform that already mixes and matches content and code from very many parties, this may make the accountability of a how an app uses ML that much harder to track.

§ 6.7.2. Possible Mitigations

Risk TA-R1 Mitigations

Note: Contributions welcome via GitHub.

Risk TA-R2 Mitigations

Note: Contributions welcome via GitHub.

Risk TA-R3 Mitigations

Note: Contributions welcome via GitHub.

Risk TA-R4 Mitigations

But they should get incentives (e.g. better performance) to use the purpose-built approach with more guarantees baked-in.

Risk TA-R5 Mitigations

Things that *end users* could be asked to do…

- permissions requests - though these are clicked away by most users.
- End users could choose to use a different model (if a browser implements a mechanism to use an alternative model, e.g. a model shipped with the browser/OS/platform locally?).

Things that *developers* could be asked to do…

- Developers could develop “model filtering” approaches, a block/accept approach for models (although places the burden on users)

Things that *implementers* could be asked to do…

- Knowing the provenance of models could help to develop a allow/block list of allowable sources for models
- Ensure / enable meaningful transparency around models, e.g. like privacy report

Things that *regulators* could be asked to do…

- Set operational requirements for characteristics of models in regulated contexts, ideally based on a neutral set of guidelines

Things that *standard makers* could be asked to do…

- Ensure Web ML guidelines are evaluatable or certifiable

Risk TA-R6 Mitigations

Wonder if something like model cards could include (or maybe they do) accountability details, or even any details at all, so that models are linked back to actual people / companies.

Risk TA-R7 Mitigations

> Note: Contributions welcome via GitHub.

§ 6.8. Sustainability

6.8.1. Risks

Risk S-R1

Web ML applications are compute / energy intensive, and widespread adoption exacerbates environmental problems.

Risk S-R2

Multiplying the value and use of ML models may create a rush to create more of them, when the environmental cost of building a model is probably high.

Risk S-R3

Distributing large ML models across the networks to each and every client may raise the environmental cost of running Web applications.

Risk S-R4

Moving ML to browsers means people have to have more powerful computers, which can be financially unachievable as well as more costly environmentally compared to a model of stronger servers and lighter clients.

Risk S-R5

ISSUE 5 Identify risk mapping for corresponding mitigation.

Risk S-R6

ISSUE 6 Identify risk mapping for corresponding mitigation.

Risk S-R7

Because inference is happening client-side, what happens to incentives for developers to make that energy efficient - i.e. if they’re not paying for the compute, do they care? It would be easy to cut corners.

Risk S-R8

Web developers have Web APIs at their disposal to help adapt the experience to be more energy efficient, see Compute Pressure API or Battery Status API. This requires balancing between enough information to satisfy the use case and not disclosing too much information to become a fingerprinting vector.

§ 6.8.2. Possible Mitigations

Risk S-R1 Mitigations

Opportunity for web browsers to make visible the energy impact of various workloads running in the browser, for example through the proposed Compute Pressure API.

Risk S-R2 Mitigations

> Note: Contributions welcome via GitHub.

Risk S-R3 Mitigations

> Note: Contributions welcome via GitHub.

Risk S-R4 Mitigations

> Note: Contributions welcome via GitHub.

Risk S-R5 Mitigations

There is probably room to improve in-browser energy impact reporting: “this tab is using significant amount of energy” – wondering if there’s room for an explicit web developer-facing API to surface energy impact more explicitly?

Risk S-R6 Mitigations

Web experiences should not depend solely on ML capabilities but enable graceful degradation path should the user or user agent wish to minimize the environmental impact.

Risk S-R7 Mitigations

> Note: Contributions welcome via GitHub.

Risk S-R8 Mitigations

> Note: Contributions welcome via GitHub.

§ 6.9. Human oversight and determination

§ 6.9.1. Risks

Risk HOD-R1

That ML models determining things like of access to welfare / insurance / healthcare etc. could rely on client-side inference?

§ 6.9.2. Possible Mitigations

Risk HOD-R1 Mitigations

> Note: Contributions welcome via GitHub.

§ 6.10. Awareness and literacy

§ 6.10.1. Risks

Risk AL-R1

The boundaries and effectiveness of ML (and its grand-sounding umbrella of artificial intelligence) may lead end users to either put more trust than they should in how well they operate, or not feel empowered to understand the impact of its use in a given web app. With the Web reaching 4bn+ users, mitigations that rely on end-users awareness are likely challenging.

Risk AL-R2

> ISSUE 7 Identify risk mapping for corresponding mitigation. ¶

Risk AL-R3

That even the designers/developers do not know of the affordances that their ML systems can provide so it can create a broader need to be able to provide feedback when something is not going well/causing harm.

Risk AL-R4

From a dev perspective: there can be an assumption that “ML will solve the problem” w/out realizing the limitations of the models/data they are employing (e.g. let’s say someone builds an app that is meant to understand facial expressions to do some action, but if people have limited facial mobility or if their models do not register their expressions as fitting into their expected classification, then the entire experience is designed around a flawed and problematic assumption that all people emote the same way).

Risk AL-R5

That without literacy and awareness users will be unable to identify the uncanny valley which can be important for privacy and security (e.g. the use of conversational bots that might be used to deceive you to gain access to your login credentials etc).

§ 6.10.2. Possible Mitigations

Risk AL-R1 Mitigations

> Note: Contributions welcome via GitHub.

Risk AL-R2 Mitigations

Perhaps specs could enable innovative use cases for developers to come up with good ways to help people be informed and aware of what’s going on under the hood with ML.

Risk AL-R3 Mitigations

> Note: Contributions welcome via GitHub.

Risk AL-R4 Mitigations

> Note: Contributions welcome via GitHub.

Risk AL-R5 Mitigations

> Note: Contributions welcome via GitHub.

§ 6.11. Multi-stakeholder and adaptive governance and collaboration

§ 6.11.1. Risks

Risk MAGC-R1

That the people who are *affected* by the outcomes of the system aren't involved in its design and development? (E.g. in a system determining eligibility for social security/benefits for people with disabilities, are people with disabilities considered as stakeholders?)

Risk MAGC-R2

"Big players" – global corporations/EU/governments – can make unilateral decisions that affect billions of people. What decision making process will they participate in?

§ 6.11.2. Possible Mitigations

Risk MAGC-R1 Mitigations

> Note: Contributions welcome via GitHub.

Risk MAGC-R2 Mitigations

It feels like the secret sauce in thinking about governance is trying to do as much as possible to build bridges across the many/various stakeholders to try to motivate maintaining and applying the principles set out in this document.

§ Appendix 1. Background: Ethics & Machine Learning

§ What is ethics?

Ethics is about what is right and wrong, good and bad. All of us think about ethics all the time as we think about what's right and wrong and make decisions about how to act accordingly.

Philosophical discussions around ethics seek to ground those thoughts about right and wrong in a rational context. They generally consider ethical issues at three levels, from the abstract to the concrete:

- **Meta-ethics** is the most abstract, concerned with questions like whether concepts of right and wrong are objective facts or subjective values
- **Normative ethics** is the more practical consideration of how we should act, both in terms of broad principles (e.g. treat others as you would want to be treated yourself) and more specific rules (e.g. do not steal)
- **Applied ethics** goes even further, considering how normative considerations should be applied in specific situations or domains, such as medical ethics, bio-ethics or AI ethics.

So ethical systems or frameworks are concerned with both broad principles to guide ethical thinking, and providing more specific answers to or guidance on a range of ethical questions.

The following is a useful sense of what ethics is/isn't from the Markkula Center for Applied Ethics

Ethics isn't:

- Legal/Corporate 'Compliance' (Legal ≠ Ethical; Ethical ≠ Legal)
- A Set of Fixed Rules to Follow (No fixed set of rules can cover all ethical cases/contexts)
- A Purely Negative Frame: ("Don't do that! Or That! Or THAT!")
- Subjective Sense of Right/Wrong ("You have your ethics, I have mine")
- Religious Belief ("It's right/wrong simply because my religion says so")
- Non-moral Customs of Etiquette ("That is just Not Done here")
- Uncritical Obedience to Authority ("Good Germans'/'Good Americans")

Ethics Is:

- Promoting objective (but context & culture-dependent) conditions of human flourishing
- Respecting the dignity of others and the duties created in our relationships to them
- Living as a person of integrity and principle
- Promoting beneficial and just outcomes, avoiding and minimizing harm to others

- Cultivating one’s own character to become increasingly more noble and excellent
- A skillful practice of moral perception, sensitivity, and flexible, discerning judgment
- Learning to more expertly see and navigate the moral world and its features

The idea of ethics as a practice is important. Ethical principles are valuable, but by themselves achieve little - they are often abstract and not directly actionable. They must be turned into concrete outcomes to have an effect. Although there may be pre-existing approaches and best practices to draw on to do this, in a fast-moving area like ML it is often necessary to think through new ethical challenges to come up with appropriate solutions.

Thankfully, applied ethics also concerns itself with the development of tools to support this type of thinking. In ML ethics, these tools help people facing ethical questions to work them through, moving from principles, to thinking about the impact of particular approaches or technologies, their benefits and potential risks and harms, and how those might be mitigated to ensure the overall ethical and beneficial impact of the approach.

This note will do the same - it will propose a set of ethical principles for Web ML, and offer guidance on how to turn those principles into practice.

For those interested to explore further:

- the Markkula Center has a useful and comprehensive set of resources devoted to technology ethics and translating principles into practice.
- The University of Helsinki has a free MOOC on Ethics of AI
- Another good MOOC is the University of Edinburgh / EdX course Data Ethics, AI and Responsible Innovation

§ 1.2 Why and how does ethics apply to machine learning?

It’s clear that ML can have a big impact on people’s lives and experience. So we could ask whether that impact is good or bad, and also how we might act to try to ensure that the impact is good rather than bad. Ethics could help us answer those questions.

Why should we take an ethical approach? Firstly, it is a deliverable required in the working group’s charter.

But as ethics is the active consideration of what’s good and bad, rather than the uncritical acceptance of rules, it’s worth considering why it should be a deliverable. There are a number of reasons:

- Technology is never neutral - it will always have social and ethical implications. The question is whether these are actively considered and addressed, or not.
- Given the scale and depth of the impact that AI/ML is anticipated to have, failure to consider the ethical implications could cause (and indeed is already causing) great harm. If technologies are not aligned with the values of the societies they operate in, they risk undermining them.
- There is clear demand for an ethical approach to ML, seen through activism from civil society, the emergence of >100 sets of ethical AI principles globally, and government moves all around the world to regulate AI.
- It aligns with the W3C's mission and design principles:
- W3C's mission is to "ensure the long-term growth of the web" - this is unlikely if web technologies create more harm than good
- W3C design principles include "Web for all" and "Web of trust" which suggest that W3C's approach to web standards is values-based
- W3C TAG is developing Ethical Web Principles. These are not normative, but provide a more explicit signal that W3C supports an ethical approach to web technologies.

§ 1.3 The universality of the web vs the specificity of ethics

The web is a universal technology, used around the world by people of all different nationalities, races, religions and beliefs.

By contrast, ethical systems are often specific to particular groups or societies - for example religions, professional groups or regions and countries.

Some beliefs may be universal across systems (e.g. a prohibition on murder), but ethical practice may also vary between different systems, cultures and contexts. Ethical principles are sometimes in tension with each other, and different societies might agree different trade-offs and balances. For example they might vary in how they balance the rights of individuals, communities and society, or security and safety and individual privacy.

Negotiating this tension is important when choosing ethical principles for something with global application like web ML standards. Several considerations are important.

Firstly, in common with other fields, ethical ML principles generally operate at a high level of abstraction, allowing them to be more universal. For example, "fairness" is a common principle, but not a specific definition of fairness, leaving that to be negotiated within any particular socio-political context.

In choosing principles, we should take an approach which supports universality. One way to do this is by choosing principles which have evidence of global relevance and support, both through the process of developing them, and their subsequent adoption.

We can also foreground principles which empower users and support their agency and autonomy, so that they can make decisions for themselves, based on their own context and values.

Also, the W3C’s role is as a promoter of open standards, rather than specific technologies or implementations. While standards are not values-neutral, in practice some of the harder ethical questions (where specific values may vary) are more relevant to specific implementations or approaches than to the broader standards. This note offers principles and guidance for implementers and authors to make use of these standards according to their context.

§ Appendix 2. Why the UNESCO principles were chosen

In response to ethical concerns and cases where ML has caused or contributed to harm, there has been an explosion in recent years of AI Ethical principles - there are now more than a hundred globally (Linking Artificial Intelligence Principles provides links to around 90 of them).

They have been developed by actors of all types, from trans-national bodies like the EU, OECD and UNESCO, to large companies, public-sector organizations, academia, private philanthropic concerns, and campaigning and activist groups.

For the W3C Web ML working group, the first question is whether we should develop our own principles from scratch or adopt some already existing ones. Given the scope of the current remit, the resources required for proper stakeholder consultation and management around developing principles from scratch, and the existence of good candidates amongst already published principles, it is proposed to adopt and adapt existing principles. This does not preclude the development of more bespoke principles from scratch in the future.

§ General considerations for choosing from existing ethical AI principles

Given the number of existing sets of principles, how should we choose amongst them?

Some key criteria are:

- They should be as universal as possible, as evidenced by:
- A diverse, global range of stakeholders involved in their development

- Broad acceptance of the final result
- They should have good coverage (be as complete as possible, while not unnecessarily broad) as evidenced by:
- Alignment with key principles found in meta-analyses of AI ethical principles, which investigate a number of sets of principles to look for convergence on common themes.
- They should align with relevant existing W3C principles and guidance in this space

§ Candidate universal ethical AI principles

For the following evaluations, see our comparison of the various principles.

Given the requirement for universality, the most likely source is transnational organizations, either governmental or non-governmental.

Some candidates:

- [UNESCO]
- [EGTAI]
- OECD AI Principles / G20 AI Principles

Of these, the UNESCO Recommendation stands out as the best candidate because:

- It is the product of an inclusive, multi-disciplinary, global consultation and development process, as part of a global institution with non-Western participants (unlike EU)
- It has been adopted by all 193 UNESCO member countries (NB: The US is not part of UNESCO and not a signatory of the new recommendations. But the UNESCO principles align with many developed in the US)
- It has a good breadth of principles (vs OECD/G20). Compared with the EU principles, it lacks an explicit statement about “Respect for Human Autonomy”. It could be argued that this is implicitly covered by the other values and principles, such as Value 1: **“**Respect, protection and promotion of human rights and fundamental freedoms and human dignity”. But while the other UNESCO Values are made more concrete by the principles that follow, Value 1 is less so, and hence there is less of an explicit commitment to Autonomy.
- Although UNESCO members are states, the principles are framed broadly enough that they can apply to anyone involved in the AI lifecycle, and the UNESCO guidance refers to ‘AI actors’ as well as states.

§ UNESCO Values and Principles

The UNESCO Recommendation consists of 4 values and 10 principles. According to the recommendation:

> Values play a powerful role as motivating ideals in shaping policy measures and legal norms. While the set of values outlined below thus inspires desirable behaviour and represents the foundations of principles, the principles unpack the values underlying them more concretely so that the values can be more easily operationalized in policy statements and actions.

The UNESCO Values are:

- Respect, protection and promotion of human rights and fundamental freedoms and human dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

The UNESCO Principles are:

- Proportionality and Do No Harm
- Safety and security
- Fairness and non-discrimination
- Sustainability
- Right to Privacy, and Data Protection
- Human oversight and determination
- Transparency and explainability
- Responsibility and accountability
- Awareness and literacy
- Multi-stakeholder and adaptive governance and collaboration

§ Sense checking UNESCO for universality

One of the main concerns in terms of universality is that many AI principles are developed from a Western perspective. So the UNESCO principles can be further sense-checked for against a set of non-Western country/region specific principles

- China Governance Principles for the New Generation Artificial Intelligence--Developing Responsible Artificial Intelligence
- Dubai AI Principles
- Japan: Social Principles of Human-centric AI

Comparing these principles (UNESCO with the ones highlighted in orange), we see good general alignment across almost all the principles and values. This is truest for the principles at their highest-level, simplest formulations. The more detailed explanation of some principles reveals particular countries' more specific policy concerns and emphases which occasionally diverge. This is not a major concern, but suggests we should exercise some caution to avoid being over-specific in fleshing out the principles.

UNESCO also continues to actively consider issues of universality and cultural diversity, and has a number of useful resources including this video on Shaping AI through Cultural Diversity.

§ Sense checking UNESCO for appropriate coverage

To check for completeness and appropriate breadth, we can compare with meta-analyses of AI Principles. There are a number of these, often referring to each other, but the following two provide a good level of coverage and depth.

- Principled Artificial Intelligence Mapping Consensus in Ethical and Rights - based Approaches to Principles for AI by Harvard Berkman Klein Centre
- The Ethics of AI: Evaluation of Guidelines by Thilo Hagendorf

Here again we can see (comparing UNESCO with the ones highlighted in pink) that there is good alignment of the UNESCO principles with the most popular principles that emerge from the meta-analyses. UNESCO covers them all, but is not going too far beyond them - the main differences are that "awareness and literacy" is not among the top principles in either meta-analysis, and "sustainability" does not appear in the Harvard one.

> RECOMMENDATION
> Given the universality and appropriate coverage of the UNESCO recommendation, it is proposed to adopt the principles and values as the basis for this document.

§ Alignment with W3C principles and values

The final consideration is how the UNESCO principles align with various relevant W3C statements of principles, values and areas of interest, and whether they need to be augmented at all.

§ Analysis of key themes from W3C documents

There are a number of different places we can look for guidance on W3C values which are relevant to ethical ML. They range from established vision to non-normative works in progress, and are reviewed in roughly that order:

W3C Vision, Mission and Design Principles

Relevant for consideration here is W3C's vision of "One web", the focus on "web for all" including accessibility and internationalization, and "web of trust", including security, privacy and trust more broadly.

These all map well to the UNESCO Recommendation. Trust is both enhanced by specific principles such as privacy and security, but also as the supporting text of the UNESCO Recommendation points out, by the effective operationalisation of the Recommendation as a whole.

W3C Horizontal Review Working/Interest Groups

Relevant here are the areas of activity of the horizontal review groups, as this is an indication that W3C places a high priority on the values they represent. The key ones are: accessibility, internationalization, privacy and security.

As above, these all map well to the UNESCO Recommendation.

W3C TAG Ethical Web Principles

Architecture is not considered relevant above as an ethical value in itself, but clearly the output of the TAG, especially their Ethical Web Principles, has strong relevance. This is a non-normative document, representing the consensus view of TAG around principles to guide their work, and that of others.

Mapping these principles against the UNESCO recommendations, we can see a good level of overlap. There are two that map less well (The web is multi-browser, multi-OS and multi-device; People should be able to render web content as they want) but this is because of their quite specific technical focus on consumption of the web, so is not considered a problem.

Less clear is "the web must enhance individuals control and power" - which seems most aligned with a principle of autonomy, but as noted above the UNESCO recommendation has only an implicit focus on that, with no explicit statement of it as a principle.

Web platform design principles

This is another TAG document which builds on the Ethical Web Principles. Relevant for consideration here are the principles in section 1: put user needs first; safety (including security and privacy and informed decision-making); trust; and meaningful consent.

Again these mostly map well to the UNESCO recommendation, with a similar note as above that informed decision-making and consent might sit most comfortably with an explicit principle of user autonomy, but can be accommodated within the other principles.

A New Focus for the W3C: Improving the Web's Integrity

According to this document, it is "intended to be a stronger vision statement for the W3C. This is currently exposed as a work item of the W3C Advisory Board, on the AB wiki" and "builds on the basis of the Technical Architecture Group's excellent Ethical Web Principles."

The values and principles include many of the themes above (accessibility, security, trust), as well as some additional ones which echo the central concerns of ethical ML (transparency, equity, fairness). They generally map well to the UNESCO Recommendation.

UNESCO's principle of "Multi-stakeholder and adaptive governance and collaboration" also aligns well with the articulation of the W3C's purpose and identity.

The area which perhaps maps least directly is W3C's concern with an interoperable, de-centralized web.

§ Summary of W3C alignment

From the above, we can see that there is generally good alignment between the UNESCO recommendation, and the values and principles which W3C has expressed in various places. There is certainly no conflict where UNESCO is proposing anything counter to W3C's values.

As noted, there are some areas where values expressed by W3C are not articulated directly as UNESCO principles, mainly things related to "Autonomy" and the concern with an interoperable, de-centralized web. The lack of Autonomy as a principle was also noted earlier in comparison with the EU Principles.

Given this …

RECOMMENDATION

It is proposed to supplement the UNESCO principles with an additional principle of "Autonomy", to make more concrete the commitment in Value 1 to "Respect, protection and promotion of human rights and fundamental freedoms and human dignity"

§ Appendix 3. Comparison of Ethics Principles

See: spreadsheet

§ Appendix 4. Workshop Format and Templates

The workshop is articulated around two documents that are expected to be completed interactively by participants:

Part One - Ethical thinking workshop - is about using the principles to generate and prioritize potential risks

Part Two - Ethical Risk Canvas - is about digging deeper into specific risks and thinking about who they might impact and how best to mitigate them.

The linked documents contain full instructions for how to use and facilitate the workshop. The format can be used in a number of ways:

- SINGLE WORKSHOP: Follow the timings for a single workshop to generate a broad overview of risks across all principles. You won’t get a lot of time to dive into mitigations for more than a couple of risks.
- MULTIPLE WORKSHOPS: If you want to spend more time thinking about risks and ensure that you identify mitigations for more risks, consider running a longer version of this workshop. For example the workshop could be run as 2 x 1h 30min sessions each around one of the activities. If you do have more time, give slightly more time to the activities and even more time to discussions. Be generous with breaks to help everyone stay fresh and engaged.
- DEEP DIVE: The activities in this workshop could also be used as a part of a longer term process by a team to thoroughly evaluate and consider the ethical risks and impacts of their project. For example, you might start by running a series of workshops to quickly gather high level thoughts about risks using the main workshop. Then, over a longer period, use workshops focused on just one or two principles to methodically build out the risks for each principle, prioritize them and explore mitigations for all prioritized risks using the Risk Canvas.

§ 7. Acknowledgments

A diverse group of experts have contributed to this document. Thanks to Ada Rose Cannon, Angel Li, Anssi Kostiainen, Christine Runnegar, Dan Appelquist, Dominique Hazael-Massieux, Hadley Beeman, Humera Minhas, James Fletcher, Janina Sajka, Jason White Joey Lee, Jonathan Bingham, Matthew Atkinson, Michael Cooper, Ningxin Hu, Rafael Cintron, Robin Berjon, Stephen Beckett, Tobie Langel, Tzviya Siegman, and many others for their feedback and comments.

§ References

§ Informative References

[AI4People]
Floridi et al. *AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*. URL: https://link.springer.com/article/10.1007/s11023-018-9482-5

[Bourtoule]
Bourtoule, L. et al. *Machine Unlearning*. URL: https://arxiv.org/pdf/1912.03817.pdf

[CDEI]
Review into Bias in Algorithmic Decision-Making. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/938857/Summary_Slide_Deck_-_CDEI_review_into_bias_in_algorithmic_decision-making.pdf

[Crawford]
Crawford, K. *The trouble with bias*. URL: https://www.youtube.com/watch?v=fMym_BKWQzk

[EGTAI]
Ethics Guidelines for Trustworthy AI. URL: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[FATML]
Principles for Accountable Algorithms. URL: https://www.fatml.org/resources/principles-for-accountable-algorithms

[Gall]
Gall, R.. *Machine Learning Explainability vs Interpretability: Two concepts that could help restore trust in AI*. URL: https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html#:~:text=Interpretability%20is%20about%20the%20extent,be%20observed%20within%20a%20system.&text=Explainability%2C%20meanwhile%2C%20is%20the%20extent,be%20explained%20in%20human%20terms.

[Leslie]
Leslie D.. *Understanding artificial intelligence ethics and safety*. URL: https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf

[Mehrabi]
Ninareh Mehrabi; et al. *A Survey on Bias and Fairness in Machine Learning*. URL:

https://arxiv.org/pdf/1908.09635.pdf

[SECURITY-PRIVACY-QUESTIONNAIRE]
Theresa O'Connor; Peter Snyder. *Self-Review Questionnaire: Security and Privacy*. 16 December 2021. NOTE. URL: https://www.w3.org/TR/security-privacy-questionnaire/

[Smuha]
Smuha, N.. *Beyond the individual: governing AI's societal harm*. URL: https://policyreview.info/pdf/policyreview-2021-3-1574.pdf

[Suresh]
Suresh. *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*. URL: https://arxiv.org/pdf/1901.10002.pdf

[UNESCO]
Recommendation on the Ethics of Artificial Intelligence. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455

[Vallor]
Vallor; Green; Raicu. *Overview of Ethics in Tech Practice*. URL: https://www.scu.edu/ethics-in-technology-practice/overview-of-ethics-in-tech-practice/

[Vaughan]
Wortman; Vaughan; Wallach. *A Human-Centered Agenda for Intelligible Machine Learning*. URL: http://www.jennwv.com/papers/intel-chapter.pdf

[Weidinger]
Weidinger et al. *Ethical and social risks of harm from Language Models*. URL: https://arxiv.org/pdf/2112.04359.pdf

[Xue]
M. Xue; et al. *Machine Learning Security: Threats, Countermeasures, and Evaluations*. URL: https://ieeexplore.ieee.org/document/9064510