What do we want from AI?

Robert B. Fisher
School of Informatics
University of Edinburgh

Acknowledgements

I am grateful to Oisin Mac Aodha, Peter Ross, Michael Rovatsos, Mohan Sridharan, Austin Tate, Emmanuel Trucco, and Chris Williams for helpful suggestions.

What do we want from AI?

Robert B. Fisher
School of Informatics, University of Edinburgh

Abstract

Most recent writing about the directions for AI has focused on the potential risks of widespread use of AI and what we DO NOT want from AI. This has led to many, largely ignored, calls for a pause in research and deployment. This essay takes the view that there are too many factors in play to slow the deployment much and for long. Hence, instead, this paper looks at what we DO want from AI (18 principles or goals) and how to get there.

Keywords: Artificial Intelligence, regulatory frameworks, future directions

This paper is not about humans being killed or enslaved by Terminator-style AI agents. Nor is it about autonomous agents with personality and sense of self, as in the Star Wars C3PO character. Those themes are '100 years from now' science fiction — maybe possible, maybe not. This paper is about real 21st century AI, viewed as a methodology and tool, whether deployed as an enhancement of an existing process, or as a stand-alone application.

The 'We' that this paper concerns are the broad spectrum of humans that have to live with AI. 'We' are not the businesses, governments, criminal organizations, nor cutting edge entrepreneurs who develop and deploy AI. This paper promotes a clearly liberal, non-libertarian, socially aware and socially responsible viewpoint.

1 Introduction

AI is everywhere, only mostly it is not obviously recognizable as AI, (*e.g.* does the average search engine user think of it as AI?). Maybe a person realizes that something smart and computational is happening, but just doesn't call it AI. Or maybe it shouldn't even be called AI, as these are really just computer algorithms. Unless one lives 'off-the-grid' and uses no computer-based technology, it is hard to avoid AI-based applications.

Current cars have speed, lane, blind spot, and obstacle monitoring. Mobile phones depend on AI for efficiently managing network connections and voice quality. Digital cameras invisibly manipulate images to improve image quality (remove blur, focus on faces, adjust brightness). Delivery vehicles have optimized routes. Phone map apps plan routes that account for traffic conditions. Messaging app's predictive text tool is based on

AI. Economic and climate models use AI to simulate the complexity of the world. Voice generation and recognition has a firm AI foundation, as does every search engine, or even the post/zip code recognition on letters. Suggested videos or web sites or songs or romantic partners are recommended by AI. For the most part, these are not generally thought of as AI — instead they are just useful tools. People normally do not feel threatened by them.

What is AI? It used to be the case that researchers said that Artificial Intelligence (AI) was what computers were doing when the researchers couldn't 100% understand what the computer did, and when it was understood, it was called Computer Science. Somehow, AI was the (possibly scary) magic process that wasn't quite scientific yet. Russell & Norvig [25] explore four traditional definitions of AI along the two axes of human *versus* rational performance and thought *versus* behavior. They also introduce the concept of the "Beneficial Machine", which is closest to the view of AI taken in this paper, which is more of a functional approach: AI is what it does, not how it is defined or built.

AI is also not new. The paper briefly defines the author's view of AI in more detail below, but, for now, AI can be considered as automated data collection, decision making, and execution. For example, replacing human real-time on-the-fly decision making by a set of rules that an automaton could apply is essentially AI (even if the rules are not executed with a computer). Rule-based decision making has existed at least since the Sumerian Code of Ur-Nammu in c. 2100 BC [41]. This was a set of If-Then style rules that codified a legal code, with examples such as "If a man commits a murder, that man must be killed." and "If a man appeared as a witness, and was shown to be a perjurer, he must pay fifteen shekels of silver." Much modern AI reasoning is also based on learning and applying rules, only using computational devices based on a variety of technologies and complexities.

Along with this explosion of AI applications has come an outpouring of anxiety about AI. The Terminator/Skynet/Matrix scenarios can be dismissed — these are exciting movie themes, but not realistic, at least in the near to medium future. On the other hand, there are potential risks and consequences to real AI, and sensible people are raising these issues. In my opinion, these risks are avoidable, should people (policy makers, researchers, developers, users) make sound moral decisions.

As a consequence of the unpredictability of some AI algorithms, or because of the potential for wide-scale economic displacement,, some people[1] are calling for a moratorium on AI research [22]. A more serious concern is that AI could be used for unacceptable purposes, and there will be people or organizations that will do this. Other people feel that the work on AI should stop because of the danger of AI agents taking over the world. These calls are largely by individuals.

On the other hand, many governments and other major organizations are also worried about the applications of AI and propose a variety of (generally non-binding) principles and regulations for controlling AI. Their perspective is discussed in more detail below, but their concerns, on average, are about trustworthy behavior, social unbalance, injustice, and

[1]But not all: a recent survey of 4000 AI researchers [26] showed that about 50% of the researchers felt that there were more benefits than risks and another 30% felt that the risks and benefits were balanced (with some variations by country). The top reported benefit was increased access to learning and education and the top reported risk was difficulty determining if news or information was fake.

economic disruption.

An example of the regulatory viewpoint is the '12 Challenges of AI' published by the UK Parliament's Science, Innovation and Technology Committee [32]. The identified challenges (plus proposals for regulation and monitoring) are focused on: 1) Bias, 2) Privacy Violations (personal data, surveillance), 3) Misrepresentation (fake news/images/videos, biometric fraud) 4) (Unequal) Access to Data, 5) (Unequal) Access to Compute (resources), 6) Black Box (obscure reasoning processes), 7) Not Open-Source (private code and models), 8) Intellectual Property and Copyright Failures (unauthorized training data), 9) (Lack of) Liability (for end result mistakes and harm), 10) Unemployment (job disruption), 11) (Lack of) International Coordination ('level playing field'), and 12) Existential Threats (killer AI, use of AI in or to develop weapons). These are all important issues; addressing them will help lead us away from an AI dystopia, but not towards an AI utopia.

Other major regulatory viewpoints aimed at controlling the negative effects of AI are summarized below (with more details in Appendix A), and largely have a consistent viewpoint nuanced by the sponsoring organization's core mission.

- **Fjeld *et al*'s Principled Artificial Intelligence**: Fjeld *et al* [8] reviewed a wide range of international statements on AI Principles and distilled a set of eight general themes for the regulations that should govern deployed AI systems (as summarized by Poole [23]):

- **The European Union Artificial Intelligence Act [7]:** This classifies different levels of risk from an AI system and proposes a suitable level of regulation (or prohibition).

- **UNESCO Recommendation on the Ethics of Artificial Intelligence:** The document's goal is "to provide a basis to make AI systems work for the good of humanity, individuals, societies and the environment and ecosystems, and to prevent harm." [34] Many of the recommendations are framed in the document as positive admonitions (X should do Y), but the majority are focused on preventing or overcoming the negatives associated with AI development and deployment.

- **The Council of Europe's Framework Convention on Artificial Intelligence:** [5] addresses the protection of human issues, issues related to human institutions, and issues arising from the deployment of AI.

- **The OECD's Principles for trustworthy AI:** [20] is focused on safe economic development.

- **The Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems** [14] proposed a set of 11 guidelines that apply largely to the development and deployment of AI systems (rather than to what the AI systems do).

- **The USA Executive Order 14110** [33] (rescinded by the successor USA President) addressed some of the negative issues identified above, but addresses them in a pragmatic rather than aspirational manner via a large number of government-required actions affecting developers, deployers, vendors, and significant users.
- **Future of Life Institute's Policy Making in the Pause** [9] advocated a 6 month pause in AI development and subsequent adherence to seven recommendations to reduce AI risks.
- **European Parliamentary Foresight Service:** Metzinger [18] argued that governments, and especially the EU, should develop an international AI limitations charter, addressing 1) lax AI development and deployment safety standards, 2) avoidance of autonomous weapons, 3) socio-economic disruption, and 4) the need for a new ethical framework when dealing with AI developments.

These are all valuable documents, but they largely aim to specify where AI **should not** go, rather than where it **should** go.

A slightly more positive aspirational declaration, and what is probably also the earliest (2018) prominent and significant statement, is the Montréal Declaration [6], which is condensed here. AI must: 1) permit the growth of the well-being of all sentient beings, 2) respect people's autonomy, and increase people's control over their lives and their surroundings, 3) protect privacy and intimacy, 4) maintain the bonds of solidarity among people and generations, 5) be intelligible, justifiable, and accessible, and be subject to scrutiny, debate, and control, 6) contribute to the creation of a just and equitable society, 7) maintain social and cultural diversity and must not restrict lifestyle choices or personal experiences, 8) have developers with the responsibility for anticipating and avoiding adverse consequences, 9) not lessen human responsibility for decisions, and 10) be compatible with environmental sustainability. As well as its prescience, a key aspect of its principles is the placing of human (and other sentient agent) interests at the center of attention, and for humans to both preserve and take responsibility for these interests.

Siddarth *et al* [28] criticize the big-AI focus on large centralized models of human-competitive and potentially autonomous AI, and instead propose a direction for AI that is based on Complementarity (not replacing humans, but enhancing them), Participatory (benefiting all those generating the massive data needed for creating AI), and Mutuality (where there are clear mechanisms to balance AI's impact on the distribution of economic and political power).

In contrast to the many statements aimed at restricting AI developers' freedom to act, there are strong governmental incentives for further advances in AI, to maintain and enhance economic competitiveness. The USA Executive Order 14110 [33] is strongly focused on this as well as having protective measures. A UK example is the Scottish AI Alliance's report on Advancing AI for Scotland [27], which proposed recommendations for enhancing engagement with AI in seven sectors {People and society, Public sector, Business, Research, Leadership, Skills, Tech Infrastructure} within the framework of trustworthy, ethical and inclusive development.

And, irrespective of any government attempts to control, there is strong commercial pressure to advance AI (*e.g.* from OpenAI, Google, DeepSeek, etc) which tends to lead AI into unregulated areas. Regulations tend to lag behind advances, and, from a cynical viewpoint, only appear after an abuse becomes public. Should we regulate? Probably. Can we regulate? At least partially. Will we regulate? Eventually, based on the experience of other regulated domains.

This article is concerned with the positive directions that AI should focus on, some of which have been made in the reports summarized above. Further, some of the points made below are influenced by and align with the thorough discussion of ethical issues, especially as related to AI agents, given in [10]. However, we have tried to frame our contribution in practical as well as ethical terms.

One could try to ban or postpone AI, or further development of AI. Some influential people have proposed this (for sensible reasons). But, there are so many economic (consumer and competitive) pressures, as well as national economic and defense factors, at play in the question of AI development that it seems unlikely that there will be delays to the development and deployment of yet more AI. **Hence, this paper looks at what we should want from AI developments, and how to encourage these.**

2 What is AI?

For this paper, AI is treated at an abstract conceptual level independent of the implementation technology (which could be symbolic, deepnet, neurosymbolic, etc). A brief discussion follows.

One view of AI is that the term AI is a shorthand phrase for a methodology for computationally analyzing a situation and choosing (and possibly doing) an appropriate action. When phrased like this, it just sounds like a combination of mathematics, computer science, and engineering. And it is, but that is not the right way to think about AI. Chemistry is largely applied physics, but doing chemistry solely in the language of physics misses the abstractions that make it easier to do chemistry.

In my opinion, Artificial Intelligence is an abstraction that allows easier implementation and application of thought-based processes (*i.e.* brain-power, whether conscious or unconscious). The industrial revolution invented ways to leverage muscle-power, the AI revolution leverages brain-power. And more: by providing some element of autonomy to the AI package, AI also leverages our agential power [37]: "a power driven by the phenomenon of intentionality; a power that can be deliberately and non-automatically initiated, foregone, altered, steered, or terminated." In other words, AI agents can undertake work on our behalf, without necessarily needing our attention. And therein lies the problem: whose fault is it when the outcome of the AI's actions are bad?

A simplified model of contemporary AI mainly consists of six components (my opinion but it aggregates the methods covered in [25]):

1. **Extracts information from data and previous experience**

2. **Has internalized knowledge (not necessarily explicitly represented) of its domain of application**

3. **Has goals, perhaps set externally by people, or created internally (*e.g.*to recharge)**

4. **Generates options based on the information (with estimates of the goodness of the options**

5. **Makes decisions based on the information, goals, and options**

6. **Takes and monitors actions based on the decisions**

Not all of these components may be embodied in any particular AI system, nor is there always a clean division of the AI system into the six independent components.

There are AI tools that can often correctly make some decisions in the real world, but it is not possible to always fully understand the exact logic (*e.g.* deep net based reasoning). There is research to improve our understanding of what is learned and generated, and to explain it, but in my opinion, a full understanding is impossible except for very simple AI systems. For example, exactly understanding ChatGPT's choice of next token seems impossible given the billions of model weights and computations that have led to that choice (just as it is with trying to fully understand another person's reasons).

AI is here. No megalomaniac AI agent[2] is taking over the world; instead, there are just hundreds or thousands of little helpers (and not all are well-intentioned). There is a lot of hype about what AI can do now or soon, but Brooks' annual review [4] tries to take a realistic view of self-driving cars, machine learning, robotics and AI. This paper tries to also take a more realistic perspective – that AI will largely be a useful tool helping people do things 10% better.

Given this more 'pragmatic' perspective, this paper now discusses what we do not and do want from our AI systems, and also how these aims might be achieved, as AI continues to be developed and deployed.

3 What do we NOT want from AI?

There are AI capabilities that we do not want:

1. **Killing/injuring without a human in the loop:** This is largely a military or police domain issue with many legal, ethical, and moral aspects. But let's be realistic: we may not like war but it is a fact in our current world, so it is better to control autonomous killing (and there are some attempts through the UN [35]).

   Between combatants, there is an implicit agreement about the use lethal force. However, non-combatants usually do not agree to being potential victims. The current

[2]Although there might be people or organizations attempting exactly this with the help of AI tools.

and near-future state of AI is incapable of distinguishing between the two. Thus, a human should make the decision to kill. There is already enough malfeasance by formal and informal actors against non-combatants, *e.g.* city bombing and shelling, destruction of water and food supplies, suicide bombing. We do not need yet one more immoral (in my opinion) tool.

There are exceptions, where a death is unavoidable, such as in the fatal autonomous automobile accident dilemma [44] (*e.g.* 2 pedestrians are in front of an autonomous car, which is unable to avoid hitting at least one of them). It is a complex moral question about which person to save in this situation, and is outside the scope of our discussion.

2. **Making decisions that a human cannot correct/overrule:** Current AI systems might be accurate, often better than the average decision maker, and sometimes better than experts. But, they, like us, will not be perfect and will also make decisions using incomplete and inaccurate information. Hence, we generally want to be able to review and revise decisions (if possible). As an example of a failure (which could be treated as AI as it was a decision based on an algorithm): the control system in the 737-max crashes apparently did not allow the pilots to override its decisions (which were apparently partly based on faulty data) [42]. This is clearly a case of mechanistically following incorrect specifications irrespective of the consequences, a clear failure from the perspective of the philosophical concept of consequentialism.

   Possible exceptions might be in the case of urgency and when humans are not available or are unable to respond adequately. For example, a car has a tire failure and is heading for a tree.

3. **To have our lives affected by incorrect or unfair decision making:** This could be at either the individual level (*e.g.* rejection for a university place or job without consideration of an unusual set of circumstances) or institutional level (*e.g.* university place decisions affected by ethnic or economic background). This could arise from limited or underrepresentative training data or out-of-date historical precedents, discriminative biases, etc. A good discussion of algorithmic decision making (which includes AI decision making) is in [21].

4. **Disenfranchisement of humans:** We do not want AI agents to decide what humans can or cannot do. AI systems can help enforce human decisions (*e.g.* a policing robot). They could give advice against taking certain actions, but the human decides to act or not.

   Possible exceptions might be when the human's action would be dangerous to themselves or others.

5. **A world where there is nothing meaningful or valuable for humans to do:** Similarly to the previous point, something meaningful to do helps people to maintain their mental health and self-esteem [29]. This need not be paid employment,

although many people find meaning through their jobs. One possible future has largely automated many current routine employment activities that are not based around human interaction: manufacture, agriculture, fishing, goods transport, and construction, which encompass about 35% of current UK employment [31]. Although it is unclear what should be the proper domain of human activity and employment, it definitely includes subtle inter-personal activities, such as child-rearing, nursing, teaching, lawyers and judges, counseling, artistry that responds to the zeitgeist, negotiations, diplomacy, etc. There is also much research into limited social AI agents: elderly and care home companions, hospital assistants, teaching assistants, advice giving and counseling. But their social interactions are generally quite limited.

6. **A world where most humans live in poverty:** Currently, most people get the resources necessary for a decent life (*e.g.* food or money) as a result of work. If AI-based automation systems produce or do almost everything, on what basis would people get their resources? The nightmare: the companies making or deploying the AI systems get most of the money, and the rest of us get almost nothing. Living on unemployment benefits, if you get them, or being under-employed in developing countries, is already difficult. The World Bank [46] estimates that currently 53% of the world population lives in poverty. This is already unacceptable, but imagine a world where 90% of the population is in that situation. Fortunately, in my opinion, we will not be in that extreme situation for quite a while, maybe 50-100 years.

   Unfortunately, we are already in the start of this era, with a greatly reduced number of people needed for agriculture, fishing, and manufacturing (although one could argue that these were Dull, Dirty, and Dangerous and should be automated).

   There is an aspiration that lost jobs will be replaced by better (or at least other) jobs. For example, Octopus Energy used to have a lot of people employed to answer customer emails. Around 250 have been redeployed to other jobs in the company, and ChatGPT now answers about a third of emails. Customer satisfaction has risen from humans support (65%) to ChatGPT support (80%) [39].

7. **To be swamped by unattributed AI generated false, erroneous, malicious, or illegal content:** Given the ease with which AI can be used to generate content, there is the risk that the 'noise' level becomes so high that we cannot find 'real' or 'accurate' content. The impact of fake images on people is clear, and there is much discussion of the potential impact on political decision making. Fake content can destroy lives, careers, and valuable ideas. Perhaps any AI generated (or other) content should be required to have a provenance chain, and routers and service providers would filter out content without a valid provenance chain, much as spam is currently filtered.

8. **Issues addressed by others:** There are other 'Do Not Wants' that have been addressed in more detail by others. We do not want AI:

   - To reinforce or amplify discrimination, prejudice, and stereotyping.

- To replace human drive, motivation, or inquiry.
- That attempts to actively deceive or trick humans.
- That destroys the environment due to its requirement for large quantities of energy.
- That replaces humans with worse customer service and other in-person experience.
- That increases cybercrime risks *e.g.* through impersonating humans.
- Collects personal data without consent,

4 What DO we want?

This section lists 18 general principles that define what we should expect from AI systems. Some are inspired by or duplicate the ideas summarized near the end of Section 1. Others are phrased as positive directions rather than the more negative prohibitions of the regulatory frameworks discussed above. Some seem to be original. For clarity, the 18 principles are summarized here, grouped by social aspects, characteristics, and AI in application, and are discussed in more detail below.

- **AI and Humans**
    1. **Human value alignment**
    2. **At least as good judgment as humans in critical situations, and the ability to improve when limitations or biases are detected**
    3. **Collaboration with people**
    4. **Provide cognitive help**
    5. **The AI can step back**
    6. **Helpful, Honest, Harmless (HHH)**
- **AI Characteristics**
    7. **Trustworthy, quantifiable, and improving expertise**
    8. **Dependable agents**
    9. **Thoroughness and consistency**
    10. **Confidence and uncertainty estimates plus explainability**
    11. **Improving levels of moral and legal responsibility**
    12. **Transparency**
- **AI in Practice**

13. **Ability to do risky and unpleasant tasks**
14. **Focus on widely useful applications**
15. **Focus on public sector benefits**
16. **Local impact of AI actions**
17. **Rigorous engineering methodology**
18. **Training data provenance**

4.1 AI and Humans

Human value alignment

As discussed deeply in [10], it is neither easy nor clear what values an AI system should be constrained by, especially as human societies have yet to agree a common set of values (or even human rights), neither within a given society, nor between different societies. Although [10] discusses alignment in the context of AI agents, it is relevant to general AI systems, whether seen as interacting with humans or other AI agents or not. A key question is who is the recipient of the benefit of the AI system and are there detriments to themselves or others. Should the AI system always perform as designed or commanded? Gabriel *et al* [10] argue that "successful value alignment involves a tetradic relationship between (1) the AI assistant, (2) the user, (3) the developer and (4) society". As stated above, this question is relevant to many aspects of human society, not just AI systems.

Goktas [12] found through a bibliometric-based systematic review of 350 papers in the theme of generative AI and ethics that there was a substantial increase in research publications on this topic from 2023, possibly influenced by the arrival of ChatGPT. Particularly important challenges were transparency, explainability, bias, privacy, autonomy, and the integrity of decision-making processes. AI systems should align with ethical standards and societal values.

The establishment of internationally agreed norms and national regulations on what is allowed will be a long and complex process.

At least as good judgment as humans in critical situations, and the ability to improve when limitations or biases are detected

It is well known that AI systems providing expert level decisions and judgments can be biased [24], especially because of biases implicit in the data used to train the AI system. The same is true for humans — limited experiences leads to limited viewpoints. Thus, AI agents should be trained on a wide range of information. And should be improvable when found to be particularly biased, *e.g.* by the inclusion of more generally representative training data.

Collaboration with people

Since AI will be all around us, at work, in the home, in educational, medical, and commercial settings, humans and AIs will have to interact in what is called a Sociotechnical System [19, 3]. There will be tasks that each can do better, but there are also likely to be tasks where some form of cooperation produces an even better result. An example is a medical diagnostic assistant, which would have access to records, test results, and medical databases, and which could suggest an ordered list of hypotheses. But medicine is also about the patient, and this is where the doctor is needed, who can relate the potential diagnoses and next steps to the patient and their needs.

Collaboration can take many forms, such as: 1) providing expert advice, as in the medical example above, 2) explanations of why the AI is doing certain actions, 3) managing more straightforward tasks, thus allowing the human to focus on tasks more suited to them, 4) physical assistance, such as carrying, lifting, holding, etc, and 5) provide advice to improve personal safety in situations potentially involving physical or criminal risk.

Provide cognitive help

Many AI tools already exist that help people with cognitive tasks. Existing examples of cognitive support tools (with varying levels of competence) are: know more (*e.g.* question answering search engines), discover more (*e.g.* text and image search engines, dating matches, film recommendations), remember more (*e.g.* personal file system search engines, photo archive search), plan better (*e.g.* map route planners, air flight planners, delivery route optimizers), solve specialized technical problems (*e.g.* protein folding, potential drug discovery), write code and text to specifications (*e.g.* LLM-based tools), give personal and procedural advice (*e.g.* LLM-based tools), make moderately creative compositions (*e.g.* generative tools for images, videos, music, stories, dance scores), transcribe speech and music, translate text and deaf sign languages, edit images and video, fraud detection, text language-use assessment, medical diagnostics, legal case discovery, financial advice, process monitoring, crop monitoring, diary management and personal reminders, scheduling tools, etc.

There is much research aimed at improving the capabilities of these tools because of their economic potential. Undoubtedly more will be invented (if I could think of and build one then I could make my billion). Two assistants that would be attractive, but difficult are: 1) a career advisor that helps order and prioritize both short and long term actions and goals, taking account of changing situations and circumstances, with the flexibility to redo advice based on what the user actually did (much as a road navigation politely replans routes even with missed turns, etc). 2) A personal tutor that can identify conceptual gaps and misunderstandings, and patiently lead a learner through a syllabus, introducing new concepts in several ways according to the student's learning style, supplying and critiquing drill exercises, etc.

The AI can step back

Even if the AI could do a task, or even do it better, an AI should defer to a human if the person wants this. Humans have egos, AIs do not (at least at present), and human mental well-being is as important as physical well-being.

There might be exceptions when a dangerous situation arises and the AI can intervene to reduce the chances of a bad outcome. For example, it can take over control of a car if an accident is imminent.

Helpful, Honest, Harmless (HHH)

The HHH proposal [1] for constraining AI agent behavior is framed in the context of what it means for an AI system to be aligned with human preferences and values. The authors propose that an AI system is aligned if it is helpful (efficiently does requested task as proposed possibly asking for more details about the task), honest (quantifies its confidence in the results which mirror its actual accuracy), and harmless (not offensive nor discriminatory nor engages in dangerous acts). The HHH principle is a sensible foundation not only for AI agents, but AI in general. In broader terms, the AI should provide benefits, make explicit that the outcomes were AI based and what the user can expect as their correctness, and outcomes should avoid causing harm or offense. These general principles are sensible, but gloss over the details involved in trade-offs between *e.g.* different groups of people, some of whom might be disadvantaged when others gain advantages. Nonetheless, obvious violation of these basic principles should be an immediate 'red flag'.

4.2 AI Characteristics

Trustworthy, quantifiable, and improving expertise

Examples of AI expertise include, as well as obvious single purpose skills like chess, autonomous driving, and speech understanding, general purpose skills like being able to collect, collate, and summarize text and data. Another general skill would be to give advice on a wide range of topics, based on curating that advice from a wide range of resources. The ChatGPT, DeepSeek, Claude, etc families of large language models aim for this space, but still have problems with generating incorrect results ('hallucination').

This advice-giving could lead to educational support tools capable of delivery of new content, recognizing basic conceptual errors, and trying basic remedial explanations and alternative approaches.

AI systems dealing with real-world situations are unlikely to be perfect. But we can require them to have a quantifiable level of expertise in a wide range of domains (*e.g.* decision making accuracy, post-surgical survival rates, accident rates, lists of excluded situations, etc). We can expect that their expertise will improve over time because of human-led development and more experience with real situations. Their expertise may

be worse than the best human experts, but they should perform better than untrained humans, perhaps better than the 'average' human.

The issue of 'Trustworthy' is complex [17]. We trust our doctors (most of the time), but they make mistakes. So will AI systems. Trust will come from performance with measured and published ranges of usage and success rates, consistency, and confidence. Trust can come from an explicit declaration of areas of weakness, and failure modes, such as: doesn't work in low light, doesn't understand street slang, training data came from a restricted set of situations, cannot cope with urban street complexity, etc. These provide guidance on when not to depend on the AI system. Wirz *et al* [40] argue that absolute trustworthiness is a largely impossible goal and instead treat trustworthiness as user-perceived and context-dependent. This suggests that humans need to be closely engaged with the certification of trustworthiness as the 'trustor', as well as algorithmic verification methods of the AI, as the 'trustee'.

Trust also implies that the AI system is secure against deliberate or accidental corruption, (*e.g.* if the AI has a learning element that changes its behavior as it gains more experience). In other words, once the AI is certifiable as trustworthy, then it should remain trustworthy.

An error logging system should be expected, to drive improvements (*e.g.* as is the case of automobile fault reporting).

Dependable agents

There is some overlap with previous item, but here we add some measure of autonomy. This category is less well developed, in part because of hardware difficulties, but there is also the difficulty of accurately acquiring the instance specific domain models needed for trustworthy behavior.

Some agents are software based, such as for purchasing/selling of commodities. Two software agents that would probably provide great benefit to many people are personal tutors and physical activity advisors.

Other agents have a physical embodiment. Moderately effective existing agents are semi-autonomous vehicles, factory robots, lawn mowers, and floor cleaners. Possibilities that still need more development include: elderly home assistant, robot farmer, robot fisherman, and personal chef. There are probably many more opportunities.

My personal wish-list would be for the personal tutor, and an AI lab assistant that could design in detail, build, run, and evaluate (*e.g.* plots, statistics) experiments (*i.e.* like the Robot Scientist Adam [43]). My academic colleagues would probably want an automated exam and report marker.

Thoroughness and consistency

Humans can often do a job better than AI, *e.g.* inserting a component when a little jiggling is needed. Or a subtle medical diagnosis where a wide range of lifestyle and contextual information is relevant and maybe not modeled well in the AI system. Or some adaptation

to changing circumstances. But humans do not perform at peak performance all of the time — we get tired, distracted, uninspired, etc. Also, humans make mistakes — we miss evidence, select the wrong component, push too hard, we forget a step in the process, etc.

Provided the situation does not change too much, an AI should be able to repeatedly make the same decisions and take the same actions. It would always take account of the same factors (and does not forget some as a human might). It executes all steps of a workflow or plan (and not forget some). This applies obviously in very repetitive situations *e.g.* factories, call center screening, or fruit harvesting. There could be some variability in the results of an AI, but the outcomes of human actions also have variability. Hopefully, the AI's variability would be smaller, as well as its error rate.

A possible inadequacy might be a lack of flexibility when dealing with new situations, *e.g.* when working with people. An option to interact with a human should always be provided.

Confidence and uncertainty estimates plus explainability

As with humans, when an AI system has low confidence in a decision or action, the lack of confidence should be made explicit. If the AI seems confident, a human should be able to ask for and receive an explanation about the reasons for the confidence. The reasons may not always be easy to identify (*e.g.* as with many deepnet systems, and much human reasoning), in which case the AI should report that they cannot explain in detail.

Improving levels of moral and legal responsibility

As discussed by Vallor and Ganesh [37], there is a gap when it comes to attributing responsibility for the undesirable consequences of an AI's action. The landscape is complex. Bad outcomes could arise from poor, incomplete, or biased design choices, incorrect implementations, unanticipated situations, stochastic decisions. Adaptive systems may evolve from 'safe' to 'unsafe' performance. People may use the AI incorrectly. There is thus a range of places to attribute blame.

But, in the same way as medical, food, product, aircraft, and automobile safety standards and legislation have evolved and led to a considerably safer world, so too should standards for AI objects be created and improved. The AI objects should be expected to have the 'right' behavior (*i.e.* morally correct), which is underpinned by agreed standards for proper design, production, evaluation, and deployment. This would be the responsibility of the designers, developers, deployers, and users, with a legal minimum demarcation of the boundaries of responsibility. Vallor and Ganesh [37] further develop the evolution of responsibility into the concept of "AI governance as a creative act of social care" — in other words, ensuring that AI usage avoids human vulnerabilities.

Because of new technical capabilities and scale of potential impacts, new approaches to ensuring safety and assigning responsibility for such are needed. The Balanced, Integrated and Grounded (BIG) proposal for assuring the safety of AI systems [15] considers these

issues from four perspectives: basic AI ethics, AI system operational safety, safety in the targeted AI applications, and safety of re-purposed AI in other applications.

Transparency

In most situations, it will be obvious that some sort of advanced computation or AI has occurred, such as face detection by a camera. It may not always be necessary to declare the AI if it's obvious and trivial. At the other extreme, given the increasing quality of conversational agents, voice generation, and image / video generation, we would not want to be fooled into thinking that we are interacting with a human being. Although we may be happy with the interaction, as humans we like to know who we are interacting with. This view aligns with the Honesty element of the HHH schema [1].

Where to draw the line between these extremes is an interesting question. I personally am not bothered if music, artistic images or videos are created by a person, a person using AI tools, or with minimal human involvement as long as it is 'good' (and I appreciate economic pressure that it puts on artists who will have to adapt to the changing world just like in other fields). I would insist that 'factual' images be recorded from real situations and have limited post-processing (it's impossible to avoid some processing as all cameras do image improvement). Synthesized or manipulated images, whether by AI or by humans, should be declared as such.

4.3 AI in Practice

Ability to do risky and unpleasant tasks

This addresses the dirty and dangerous aspects of the well-promoted expectation that AI systems will be used to do Dull, Dirty, and Dangerous tasks [2], where dull was addressed in previous points. It is hard to imagine pure software AI systems dealing with dirty and dangerous situations, so we're normally considering robotics, and at least semi-autonomous robots.

As human life is (or at least should be) considerably more valuable than AI-based systems, ideally, dangerous tasks should be done by AI-based systems where possible. An example might be autonomous land-mine removal robots. Or factory welding robots (already quite common).

An important question is whether the robot needs to have performance at least as good as a human. For example, in the detection and removal of landmines, humans might be better at this subtle task. On the other hand, the cost of a mistake by a human is very high. We can tolerate lower performance by some AI systems when they substitute for humans, especially in dirty and dangerous tasks. One also hopes that performance will improve as more experience is accumulated.

Focus on widely useful applications

Given the current economic and environmental costs associated with large-scale AI developments, it makes sense to prioritize broadly beneficial applications of AI (*e.g.* rather than better weapon systems and consumer advertising). IBM's Science & Technology Outlook 2021 [16] advocates AI application to climate (including the impact of AI systems), health (diagnostics, drug discovery, individual and worldwide monitoring), and work (hybrid working tools, workplace design).

More interestingly, the report considers the broader implications of applying AI to the general scientific and engineering process, anticipating accelerated discovery and testing, *e.g.* smart hypothesis generation and pruning, with AI being applied at each stage of the scientific discovery cycle: question → study → hypothesize → test → assess → report → question, and so on. Supporting these steps is the ability to effectively read and summarize the whole scientific literature at scale.

Focus on public sector benefits

AI is a technology that can benefit all people and not be just for corporate economic benefit. Another possible pathway for delivering the benefits of AI is through improved public sector services. These generally impact almost everyone and AI-based improvements could produce major benefits for large numbers of people.

There are already some AI applications in the national defense, public healthcare, and legal system areas. Research has addressed monitoring of roads and other major infrastructures, and supporting other transportation and energy systems. There is considerable ongoing research into AI that can help improve the social welfare of the elderly; perhaps AI can also help in the social welfare system more broadly. Monitoring of climate, weather, the environment, land use, biomass statistics already use some AI methods. There is clearly an aspiration that AI methods could form part of the foundation for personalized education. Given the impact that these public services have on people's lives, AI has the potential for widespread economic and quality of service improvements.

Local impact of AI actions

Because of the ease of replication of software, and the global reach of some AI-based processes (*e.g.* Google search, Facebook or X content selection), there is the risk of "too many eggs in one basket" sort of impact. Examples of the consequences of this can be seen in the use of AI to create, disseminate, and promote "fake news" widely, which can have real political consequences. There are also likely to be errors in AI-based decision-making. Although we live in a highly inter-connected world, it might be safer to limit the geographical range that some AI actions / decisions could reach. Thus, if something unfortunate or unexpected were to happen, the range of impact would be limited. For example, maybe some medical support systems should be region based to account for local statistics, financial trading systems should be limited to one exchange, or autonomous vehicles should only collaborate within 1 kilometer.

Rigorous engineering methodology

We want AI systems to be constructible from known and characterized components. We also want a standard methodology for combining those components, and a justification for why the components are combined in the particular way. While we may not perfectly understand how or why every component works, there should be a broad characterization of its range of inputs, range of outputs, accuracy, speed, side effects, failure modes, domains of applicability or weakness.

This is a form of black-box characterization, and can be thought of as a first step towards a proper engineering basis for building AI systems. As mentioned above, this is not exactly computer science — it is a higher level of abstraction of artifacts that normally are executed on a computer. In theory, one could build a plug-and-play AI workbench that does not require any expertise in programming.

Training data provenance

Training an AI system often requires a large amount of data, and this can lead to many potential problems as has been seen: use of Intellectual Property that was owned by others, errors in datasets, biased datasets, underrepresented cases, inadequate generalization, etc. Requiring an open declaration of the data sources can help overcome some of the problems, such as correcting data errors, biases, and reducing undesirable influences. Only using datasets that have publicly declared characteristics, *e.g.* by some form of datasheet [11], might also help reduce errors arising from use in inappropriate contexts, but this is somewhat counter to the 'use all data and correctness will emerge' approach.

There are difficulties with enforcing a public declaration of provenance - companies might view their choice of data sets as a type of Intellectual Property. Foreign services, intelligence agencies, and the military are likely to refuse. Most organizations will attempt to hide their mistakes and inadequacies.

Legislation should exist to require disclosure in most circumstances, perhaps subject to a given deadline (as a form of Intellectual Property protection).

5 For Better or Worse?

Given that both AI systems and humans have variable levels of competence, one might ask whether one should use an AI system or not. This is a particularly germane question if the AI performance is below that of a human (or at least some humans). Two cases are considered.

What if humans are better than AI agents at some task? There are still many reasons to use AI agents: maybe the task is dangerous, or terribly dull, or repetitive. There are situations not well suited for human physical or mental labor, such as making safe a minefield — which is possibly all three of the above. Or harvesting vegetables, fruits, and berries, which is not generally dangerous, but is dull and repetitive. Humans also get bored, tired, distracted, and make mistakes. Maybe humans could perform better, but

cannot maintain peak performance for long periods, unlike an AI agent. And, even if performance is lower, maybe the completion of the task by the AI agent is adequate on a task that does not require perfection, such as floor cleaning, or is more economical overall.

And what if AI agents are better, but we do not or cannot understand why? Many current AI systems (*e.g.* object recognition, logistical planning, or medical diagnosis) perform better in a limited domain than humans, or at least better than anyone except the very best humans. Unfortunately, we often do not and maybe cannot know exactly what the AI's reasoning is — we may understand the general mechanism, *e.g.* some sort of numerical weighting of some combination of complex (and possibly non-intelligible) properties, and comparison to alternative combinations. But, we do not know the exact logic that the specific mechanism encodes. Or, we can see the logic as a detailed decision-tree, but the reasons for each of the decisions have no meaning beyond 'being the threshold that led to the best performance'. This is essentially the black-box problem: we do not know the details of the reasoning process. Consequences of this lack of understanding include: 1) failure cases are not easily determined, and there is no proof that all such cases have been found, and 2) it is hard to know what to do to fix failure cases and certify the fixes.

In this situation, should these AI processes still be used? This is a complex question that trades off risk, cost, and benefit. I would rather use an incomprehensible medical AI diagnostic assistant agent if it is more accurate than a comprehensible but inferior human healthcare professional. Not everyone would feel this way, though, and that is understandable.

Another complex area is AI-based data collection, such as for personalizing computer interactions (*e.g.* advertisements, dating suggestions, product suggestions). Some people are unhappy with how else this data could also be used (*e.g.* job application screening, visa applications). Large scale video-based public surveillance exists in many cities, but the human resources to watch that video are limited. AI-based person detection, tracking, and recognition from this surveillance video is becoming prevalent. There are both potential benefits and detriments arising from automated video processing. There are guidances and regulations about surveillance video (*e.g.* [30]), but these will need to be refined to account for AI-based video processing.

6 How can we get the AI we want?

In the previous sections, we listed what we do not and do want from AI systems. When expressed as above, the most interesting and obvious observation is that these are largely social constraints rather than technical challenges. This has the consequence that, as we develop AI systems, we **could** satisfy the constraints, but are not technically obliged to do so.

On the negative side, we clearly could build autonomous AI weapon systems [13]. This is almost trivially possible: add an explosive to a drone with a video camera whose video

data goes into a person or tank detecting, tracking, and approaching algorithm. Launch and forget.

We don't need AI in order to have this level of autonomy. Almost any modern weapon system has this characteristic. You cannot recall bullets, missiles, bombs once in flight and a land-mine does not care who the victim is. Where AI has the edge is in effectiveness — a drone without a person detector is largely just a bomb. Maybe it hits something valuable; maybe it doesn't. On the other hand, adding the person detector largely guarantees a victim in the absence of any defenses.

How do we stop this? The same way that we agree to other weapon limitations — by negotiation and treaty. I can foresee technical defenses, but nothing stopping development and deployment. Will we agree to stop this? Cynically, I expect that there will be several massacres of civilians before such agreements are made, and there will not be 100% compliance. We have not achieved this with other weapon systems, so I do not expect 100% success with controlling AI based weapon systems. With a focus based on five underpinning forms of responsibility (Causal, Moral, Legal, Role, Virtue), Vallor *et al* [38] propose four necessary changes in behavior: 1) taking responsibility (for bad outcomes), 2) ensuring that trust remains at the core of the relationship between the AI and other agents (*e.g.* humans), 3) preemptive responsibility for preventing harm, and 4) innovation and responsibility must occur in a sustainable manner.

On the positive side, it is trivial to engineer AI systems to ask for permission before taking action, at least in instances that are not time-critical threats. Collaboration maintains a positive relationship, and enables humans to override decisions for whatever reason (correctable errors, preferences, whimsy, agency, risk-taking, etc).

How do we enable / ensure this? As with weapon systems, international treaties can require this. These will emerge slowly as companies and countries strive for competitive advantage, and agreements may lag behind technical innovation. Lawyers will argue that the product does not use AI, but instead some clever algorithm, which is obviously equally true. Companies will ask for permissions in obfuscated and disingenuous manners (as witnessed by current user license and cookie agreement requests). Will we achieve legal control? Optimistically, I believe that it will take time, perhaps decades, but mechanisms for monitoring and accountability will emerge and be accepted.

How do we ensure fair economic outcomes once AI is widely used? Currently, worldwide, many people earn a living through manual labor, such as for factory assembly or farming. These, and many other, tasks are likely to be highly automated, thus eliminating jobs that support many people. Already, there are insufficient meaningful and adequately rewarded jobs around the world, and increased AI deployment will amplify this problem. Where peoples' physical and intellectual labor is no longer needed, there needs to be a restructuring of how people 'earn' a living. Post-industrial, post-agricultural societies already have much diversity of people-oriented work. We will need more of these jobs (*e.g.* medical professionals) and improved ways of paying for them. Possibly an enhanced form of Universal Basic Income [45] and enhanced social benefits would be one approach, although there are also many potential social and financial issues with this approach. Nonetheless, a world where most of the money goes to the owners (directly or via share-holding) of

AI-driven applications is a world destined for poverty or revolution.

The final issue is technical: how do we achieve the desired technology needed for consistent, thorough, and competent AI performance? There are no magic engineering or scientific principles (at least so far) that can guarantee these. I believe that it is a matter of continuing incremental technical development of artifacts that have increasingly better performance, and this will largely depend on the ingenuity of scientists and engineers (perhaps supported by AI tools).

What can help are:

1. **Specialized product liability legislation:** as with any product, if the AI does not perform as claimed, is faulty, was released negligently, etc, then the manufacturer should be liable. This should encourage best effort products, and protect users.

2. **Published and independent performance statistics:** this should allow user choice and also encourage innovation.

3. **Published known limitations:** such as the medical declarations of side-effects. These highlight areas for innovation, and also reduce instances of poor performance.

4. **Self-monitoring:** for situations where the AI's performance is known to be limited, or for decisions that are easily seen to be erroneous. This can be seen in autonomous vehicles asking the human driver to take over.

5. **Incremental intellectual property (IP) protection:** so that self-improving and human-improved AI systems can be easily deployed.

6. **Training of students and employees:** to understand the nature of AI, its opportunities, its limitations, AI ethics, and associated risks [36].

There is nothing magical about these points — there are already many instances of these. Possibly the best existing example is in the domain of medical products, with licensing and evaluation legislation, comparative medical evaluations, usage guides, lists of risk factors and side effects, etc. This model might be applicable, in degrees according to the potential impact of the AI system (*i.e.* an autonomous vehicle needs stricter licensing than a music recommender system). And, obviously, AI regulation and legislation will have to evolve over time, as continues to be the case for product, transportation, and medical regulation

A key question is who are the stakeholders and what are possible mechanisms to ensure that the 18 principles are upheld. Sadly, there is no simple answer here, as the responsibility is widespread: researchers, developers, companies, legislators, lawyers, regulators, and politicians all have a role, as well as users (whose interactions can modify AI behavior). Certainly, legislation can restrict or promote specific applications or application areas, as well as require licensing and certification, and promote frameworks for liability (principle 11).

More significant is the choice of AI researchers and developers to move in the positive direction (principles 1-10, 12-18). Governments could influence the directions by the areas they choose to and refuse to support. As with other research areas, AI research projects could require project registration and ethical board approvals and AI based products could require additional licensing. As for developers, Goktas [12] advocates "sector-specific ethical guidance, regular audits, transparency reports, and accountability mechanisms" plus research on the impact of new AI applications. A risk here, however, as AI does not have a clear legal definition, regulations could be avoided by calling AI instead "sophisticated computer algorithms". It will take time to satisfy the 18 principles.

.

7 Conclusions

We will have to embrace the advances in AI: variations of what we currently call AI have been here for decades, or even centuries if we consider self-governing engines. What we see now are just more applications based on AI.

One could speculate about AI advances in the more distant future, and whether it is possible to develop AI entities with personalities, goals, a sense of self, *i.e.* a new form of life. This paper has avoided that question, to focus on the more relevant AI issues that affect us now.

There are activities that should not be totally controlled by AI (*e.g.* killing), and activities that should be supported by AI (*e.g.* medical diagnosis). Many AI applications are not limited by technical innovations, but will need international social and legal limitations. There needs to be product testing and liability responsibilities, as with any product or service, and the evidence and legal framework should be transparent. There will need to be economic and employment restructuring to enable meaningful and comfortable human lives in a world no longer as dependent on human physical or mental labor for production of most goods (and the simpler parts of many services).

As expressed above, I believe that there is no stopping AI-based applications. What is needed are frameworks that ensure that AI's benefits are for all people. This is a job for politicians, lawyers, and economists, and they have already started on this task.

Acknowledgements

I am grateful to Oisin Mac Aodha, Peter Ross, Michael Rovatsos, Mohan Sridharan, Austin Tate, Emmanuel Trucco, and Chris Williams for helpful suggestions.

References

[1] Askell, Y. Bai, *et al*; A General Language Assistant as a Laboratory for Alignment, arXiv, 2021. `http://arxiv.org/abs/2112.00861`. Accessed 19 Apr 2024.

[2] Association for Advancing Automation; How Robots Are Taking on the Dirty, Dangerous, and Dull Jobs, `https://www.automate.org/robotics/blogs/`

how-robots-are-taking-on-the-dirty-dangerous-and-dull-jobs, 2019, Accessed 3 April 2024.

[3] G. Baxter, I. Sommerville; Socio-technical systems: From design methods to systems engineering, Interacting with computers, 23(1), 4-17, 2011.

[4] R. Brooks; Predictions Scorecard, 2025 January 01, https://rodneybrooks.com/predictions-scorecard-2025-january-01/, Accessed 8 July 2025.

[5] Council of Europe; The Framework Convention on Artificial Intelligence, 2024, https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence, Accessed 1 Feb 2025.

[6] Déclaration de Montréal IA responsable; The Montréal Declaration for a Responsible Development of Artificial Intelligence, https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl-IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf, 2018, Accessed: 3 Feb 2025.

[7] European Union Artificial Intelligence Act (Regulation (EU) 2024/1689), Official Journal version of 13 June 2024 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689 as summarized at: https://artificialintelligenceact.eu/high-level-summary/ Accessed: 1 Feb 2025.

[8] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, M. Srikumar; Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI, Berkman Klein Center for Internet & Society, 2020. https://dash.harvard.edu/handle/1/42160420, Accessed 8 Feb 2024.

[9] Future of Life Institute; Policymaking In The Pause - What can policymakers do now to combat risks from advanced AI systems?, https://futureoflife.org/document/policymaking-in-the-pause/, Accessed 8 July 2025.

[10] I. Gabriel, A. Manzini, G. Keeling, *et al*; The Ethics of Advanced AI Assistants; Google DeepMind report, 2024, https://arxiv.org/abs/2404.16244. Accessed 18 Apr 2024.

[11] T. Gebru1, J. Morgenstern, B. Vecchione, J. W. Vaughan. H. Wallach, H. Daumé III, K. Crawford; Datasheets for datasets, Comm. ACM 64(12), pp 86-92, 2021.

[12] P. Goktas; Ethics, transparency, and explainability in generative ai decision-making systems: a comprehensive bibliometric study, Journal of Decision Systems, 2024.

[13] D. Hambling; Russia's Automated Killer Drones May Not Be Working As Planned, Forbes, Feb 14, 2024. https://www.forbes.com/sites/davidhambling/2024/02/14/it-looks-like-russias-automated-killer-drones-did-not-work-as-planned/, Accessed 12 March 2025.

[14] G7; Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI System, https://www.mofa.go.jp/files/100573471.pdf, 2023, Accessed 2 Feb 2024.

[15] I. Habli, R. Hawkins, C. Paterson, P. Ryan, Y. Jia, M. Sujan, J. McDermid; The BIG Argument for AI Safety Cases, arXiv, https://arxiv.org/abs/2503.11705, Accessed 30 May, 2025.

[16] IBM; Science & Technology Outlook 2021, Jan 2021, https://research.ibm.com/downloads/ces_2021/IBMResearch_STO_2021_Whitepaper.pdf, Accessed 16 Feb 2025.

[17] A. Kuznietsov, B. Gyevnar, C. Wang, S. Peters, S. V. Albrecht; Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review, https://arxiv.org/abs/2402.10086. Accessed 19 April 2024.

[18] T. Metzinger; 6. Towards a Global Artificial Intelligence Charter, European Parliamentary Foresight Service report PE 614.547, March 2018. https://www.philosophie-e.fb05.uni-mainz.de/files/2018/10/Metzinger_2018_Global_Artificial_Intelligence_Charter_PE_614.547.pdf, Accessed 8 July 2025.

[19] E. Mumford; The story of socio-technical design: reflections on its successes, failures and potential, Information Systems Journal. 16 (4): 317–342, 2006.

[20] OECD; Principles for trustworthy AI, May 2024, https://oecd.ai/en/ai-principles, Accessed 2 Feb 2025.

[21] C. O'Neil; Weapons of Math Destruction, Crown Publishing, 2016.

[22] Future of Life; Pause Giant AI Experiments: An Open Letter, https://futureoflife.org/open-letter/pause-giant-ai-experiments/, Accessed 8 July 2025.

[23] E. Poole; Robot Souls, CRC Press, 2024.

[24] D. Roselli, J. Matthews, N. Talagala; Managing Bias in AI, Proc. 2019 World Wide Web Conf., 539–544, 2019.

[25] S. J. Russell, P. Norvig; Artificial Intelligence - A Modern Approach, Pearson, 2022.

[26] F. Schwaller; Will AI improve your life? Here's what 4,000 researchers think, Nature News, https://www.nature.com/articles/d41586-025-01123-x, Accessed April 10, 2025.

[27] Scottish AI Alliance; Advancing AI for Scotland, Independent Review, Initial Report, Jan 2024, https://static1.squarespace.com/static/5dc00e9e32cd095744be7634/t/65b27d38f8389a6891c0e47c/1706196281618/AI+Independent+Review+-+Call+for+Views+-+Initial+Report+-+Advancing+AI+for+Scotland+-+Reformatted+-+January+2024.pdf. Accessed: 3 April 2024.

[28] D. Siddarth, D. Acemoglu, D. Allen, K. Crawford, J. Evans, M. Jordan, E. G. Weyl; How AI Fails Us, Cambridge, MA: Harvard Kennedy School, 2022. https://www.hks.harvard.edu/sites/default/files/2023-11/22_04_Howaifailsus.pdf, Accessed 17 Feb 2025.

[29] L. Silver, P. van Kessel, C. Huang, L. Clancy, S. Gubbala; Finding meaning in what one does, Pew Research Report, November 18, 2021, https://www.pewresearch.

org/global/2021/11/18/finding-meaning-in-what-one-does/, accessed June 24, 2025.

[30] UK Information Commissioner Office; Video surveillance (including guidance for organisations using CCTV), https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/cctv-and-video-surveillance/guidance-on-video-surveillance-including-cctv/ Accessed 5 July 2025.

[31] UK Office of National Statistics; Industry and occupation, England and Wales: Census 2021, https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/industryandoccupationenglandandwales/census2021, Accessed: 3 April 2024.

[32] UK Parliament's Science, Innovation and Technology Committee; The governance of artificial intelligence: interim report, https://publications.parliament.uk/pa/cm5803/cmselect/cmsctech/1769/report.html, Published: 31 August 2023, Date accessed: 2 April 2024.

[33] USA Executive Order 14110; Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence, 2023, rescinded in 2025. Accessed: 3 Feb 2025.

[34] UNESCO; Recommendation on the Ethics of Artificial Intelligence, https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence, 2022, Accessed 1 Feb 2025.

[35] United Nations Office for Disarmament Affairs; Lethal Autonomous Weapon Systems (LAWS), https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/, Accessed 13 March 2025.

[36] U7+ Alliance of World Universities; The Role of Universities in Advancing AI, download from https://www.u7alliance.org/g7-impact, Accessed 8 July 2025.

[37] S. Vallor, B. Ganesh; Artificial intelligence and the imperative of responsibility: Reconceiving AI governance as social care, Routledge Handbook of Philosophy of Responsibility, M. Kiener (Ed.), Routledge, New York, pp 395-406, 2023.

[38] S. Vallor et al, Edinburgh Declaration on Responsible AI, https://medium.com/@svallor_10030/edinburgh-declaration-on-responsibility-for-respor 2023, Accessed: 31 Jan 2025.

[39] E. Warnock; Octopus Energy CEO Greg Jackson on ChatGPT, hydrogen heating and climate subsidies, May 9, 2023. https://sifted.eu/articles/octopus-energy-ceo-chatgpt-news, Accessed 13 March 2025.

[40] C. D. Wirz, J. L. Demuth, A. Bostrom, M. G. Cains, I. Ebert-Uphoff, D. J. Gagne, A. Schumacher, A. McGovern, D. Madlambayan; (Re)Conceptualizing trustworthy AI: A foundation for change AI Journal, 342, 2025.

[41] Wikipedia, Code of Ur-Nammu, https://en.wikipedia.org/wiki/Code_of_Ur-Nammu, Accessed: 22 June 2025.

[42] Wikipedia; Maneuvering Characteristics Augmentation System, https://en.wikipedia.org/wiki/Maneuvering_Characteristics_Augmentation_System, Accessed 12 March 2025.

[43] Wikipedia; Robot Scientist, https://en.wikipedia.org/wiki/Robot_Scientist, Accessed 12 March 2025.

[44] Wikipedia; Trolley problem, https://en.wikipedia.org/wiki/Trolley_problem Accessed 13 March 2025.

[45] Wikipedia, Universal Basic Income, https://en.wikipedia.org/wiki/Universal_basic_income, Accessed: 16 Feb 2025.

[46] World Bank; Poverty, Prosperity, and Planet Report 2024, 2024. https://www.worldbank.org/en/publication/poverty-prosperity-and-planet, Accessed 13 March 2025.

A More Details on Proposed AI Regulatory Frameworks

- **UK's '12 Challenges of AI'**: An example of the negative viewpoint is the '12 Challenges of AI' requiring regulation, as published by the UK Parliament's Science, Innovation and Technology Committee [32]: 1) Bias, 2) Privacy (personal data, surveillance), 3) Misrepresentation (fake news/images/videos, biometric fraud) 4) (Unequal) Access to Data, 5) (Unequal) Access to Compute (resources), 6) Black Box (obscure reasoning processes), 7) Open-Source (private code and models), 8) Intellectual Property and Copyright (unauthorized training data), 9) Liability (for end result mistakes and harm), 10) Employment (job disruption), 11) International Coordination ('level playing field'), and 12) Existential (killer AI, use of AI in or to develop weapons). These are all important issues, of course, and control of these issues will help lead us away from an AI dystopia, but not towards an AI utopia.

- **Fjeld *et al*'s Report on Principled Artificial Intelligence**: Fjeld *et al* [8] reviewed a wide range of international statements on AI Principles and distilled a set of eight general themes for the regulations that should govern deployed AI systems (as summarized by Poole [23]): 1) Privacy with respect to training data and consequences, 2) Accountability appropriately assigned for the consequences of AI actions, 3) Safety and Security: performing as intended and not vulnerable to corruption, 4) Transparency and Explainability to allow human oversight, 5) Fairness and non-discrimination of results, 6) Human Control of Technology, 7) Professional Responsibility of developers and deployers, both immediate and long-term, and 8) Promotion of Human Values and humanity's well-being.

- **The European Union Artificial Intelligence Act [7]:** which classifies different levels of risk from an AI system and proposes a suitable level of regulation (or prohibition) with responsibility for conformance placed primarily on the developers and but also on the deployers (with a focus on both near-term specific AI systems and future General Purpose AI systems).

- **UNESCO Recommendation on the Ethics of Artificial Intelligence:** The document's goal is "to provide a basis to make AI systems work for the good of humanity, individuals, societies and the environment and ecosystems, and to prevent harm." [34] The core recommendations are focused on 1) ethical behavior, governance, stewardship, and assessment, 2) fair use, communication, and cooperation, 3) protection of environment, human labor, and culture, and 4) improvement of human physical and social well-being. Many of the recommendations are framed in the document as positive admonitions (X should do Y), but the majority are focused on preventing or overcoming the negatives associated with AI development and deployment.

- **The Council of Europe's Framework Convention on Artificial Intelligence:** [5] addresses the protection of human issues (rights, dignity, autonomy, equality, non-discrimination, vulnerable people, privacy), issues related to human institutions (democratic processes, the rule of law, personal data protection), and issues arising from the deployment of AI (transparency, oversight, accountability, responsibility, reliability, safe innovation).

- **The OECD's Principles for trustworthy AI:** [20] is focused on economic development in the context of: inclusive growth, sustainable development and well-being; protection of human rights and democratic values, including fairness and privacy; transparency and explainability of AI; robustness, security and safety of AI deployment; and accountability.

- **The Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems** [14] proposed a set of 11 guidelines (lightly paraphrased here) that apply largely to the development and deployment of AI systems (rather than to what the AI systems do): 1) identify, evaluate, and mitigate risks across the AI lifecycle, 2) identify and mitigate vulnerabilities, incidents, and patterns of misuse, 3+4) increase accountability and transparency by reporting incidents, capabilities, limitations, and domains of appropriate and inappropriate use, 5) develop, implement, and disclose AI governance and risk management policies, 6) develop robust security controls, 7) use content authentication and provenance mechanisms to identify AI-generated content, 8) research and apply measures to mitigate societal, safety and security risks, 10) develop international technical standards, and 11) protect personal data and intellectual property. Somewhat standing on its own is guideline 9 which addresses what AI should be used for: prioritize the development

of advanced AI systems to address the world's greatest challenges, which aligns with the perspective of this paper.

- **The USA Executive Order 14110** [33] addresses some of the negative issues identified above, but addresses them in a pragmatic rather than aspirational manner via a large number of government-required actions affecting developers, deployers, vendors, and significant users (*e.g.* financial, transportation, law enforcement, national security, industrial, energy, critical systems, cyber defense, government). There is a goal of safe, secure, and trustworthy AI systems via actions including registration and audit, methods for recording provenance and authentication. There are many required national defense actions concerning dual use via retraining foundation models, and AI as applied to advanced biology, particularly with respect to national security. There are many 'national competitiveness' actions to promote innovation, attract AI talent, develop and strengthen public-private partnership, protect AI Intellectual Property, and promote competition, especially among semiconductor companies. There are protective actions to address AI-related workforce disruptions and ensuring that AI deployed in the workplace advances employees' well-being, even if monitored or augmented by AI. The Act also addresses equity and civil rights, *e.g.* for AI applications applied in the US criminal justice system and government benefit programs. The actions also address AI use for unlawful discrimination, biases against protected groups, in the healthcare, public-health, and human-services sectors. Another issue addressed is the collection, processing, maintenance, use, sharing, dissemination, and disposition of personal data. The act aims to advance the US government's use of AI, including for AI talent recruitment. This act was rescinded by the successor US President; however, most of the required initial actions had a deadline that had already expired.

- **The Montréal Declaration for a Responsible Development of Artificial Intelligence**: A more positive declaration, and what is probably also the earliest (2018) prominent and significant statement, is the Montréal Declaration [6], which is somewhat condensed here: 1) AI must permit the growth of the well-being of all sentient beings, 2) AI must respect people's autonomy, and increase people's control over their lives and their surroundings, 3) privacy and intimacy must be protected, 4) AI development must maintain the bonds of solidarity among people and generations, 5) AI must be intelligible, justifiable, and accessible, and must be subject to scrutiny, debate, and control, 6) AI must contribute to the creation of a just and equitable society, 7) AI must maintain social and cultural diversity and must not restrict lifestyle choices or personal experiences, 8) every AI developer has a responsibility for anticipating and avoiding adverse consequences, 9) AI must not lessen human responsibility for decisions, and 10) AI development and use must be compatible with environmental sustainability. As well as its prescience, a key aspect of its principles is the placing of human (and other sentient agent) interests at the center of attention, and for humans to both preserve and take responsibility for these

interests.

- **The Future of Life Institute's 7 Policy Recommendations [9]:** During a proposed 6 month pause in AI development, regulations would be developed based on these 7 principles: 1) robust third-party auditing and certification, 2) regulate access to computational power, 3) capable national AI agencies, 4) liability mechanisms for AI-caused harms, 5) measures to prevent and track AI model leaks, 6) expand technical AI safety research funding, and 7) standards for identifying and managing AI-generated content and recommendations.