24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues

Jose Hernández-Orallo[1] and **Fernando Martínez-Plumed**[1,2]
and **Shahar Avin**[3] and **Jess Whittlestone**[3] and **Seán Ó hÉigeartaigh**[3]

Abstract. AI safety often analyses a risk or safety issue, such as interruptibility, under a particular AI paradigm, such as reinforcement learning. But what is an AI paradigm and how does it affect the understanding and implications of the safety issue? Is AI safety research covering the most representative paradigms and the right combinations of paradigms with safety issues? Will current research directions in AI safety be able to anticipate more capable and powerful systems yet to come? In this paper we analyse these questions, introducing a distinction between two types of paradigms in AI: artefacts and techniques. We then use experimental data of research and media documents from AI Topics, an official publication of the AAAI, to examine how safety research is distributed across artefacts and techniques. We observe that AI safety research is not sufficiently anticipatory, and is heavily weighted towards certain research paradigms. We identify a need for AI safety to be more explicit about the artefacts and techniques for which a particular issue may be applicable, in order to identify gaps and cover a broader range of issues.

1 INTRODUCTION

As in any other scientific or engineering discipline, many AI researchers work within a well-established paradigm, with some standard objects of study, problems to solve, and associated formalisms and terminology. Understanding past, current and future AI paradigms is an important source of insight for funding agencies, policymakers, and AI researchers themselves, because it shapes how we think about what problems AI research is aiming to solve, what methods are required to solve them, and what the wider implications of progress might be.

We believe that thinking clearly about AI paradigms is particularly crucial for *AI safety* research: an increasingly important area concerned with understanding and preventing possible risks and harmful impacts in the design and deployment of AI systems. For the purposes of this paper, we define AI safety broadly, to include both risks from AI systems for which the source of risk is accidental, ranging from unpredictable systems to negligent use, and non-accidental risks such as those stemming from malicious use or adversarial attacks (which might sometimes also be referred to as AI security risks.) This includes risks with many different types of consequences: including human, environmental, and economic consequences. AI safety is becoming particularly important as AI is increasingly used to automate tasks that involve interaction with the world. A characteristic example is AI being used in many components of self-driving cars such as perception, reasoning and action.

Research in AI safety has typically analysed a specific risk or issue under a particular existing paradigm, such as interruptibility for reinforcement learning agents [33], adversarial attacks on deep learning systems [43], or fake media produced with GANs [21]. Poor choices or misinterpretation of current and future paradigms may result in AI safety research focusing in the wrong places: analysing scenarios that will not take place in the future, or that will at best manifest in completely different ways. For example, concerns about adversarial attacks on current deep learning methods have led to many papers proposing methods to defend against malicious perturbations, but it is not clear that these actually relate to plausible security concerns [16]. Though adversarial examples help us understand how brittle current deep learning methods can be, the kinds of attacks that some safety research is concerned with can only arise in contrived interactions, and make strong assumptions about the goal, knowledge, and action space of the attacker. In contrast, other risks have been ignored because they are outside of the current paradigm: for instance, attacks which go beyond an "independent" or functional interpretation of a neural network, and instead use the "information from previous frames to generate perturbations on later frames" [41].

Thinking about different paradigms is also important for clearly assessing safety considerations and risks in concrete real-world applications. For instance, the risks of using AI in vehicles will be perceived differently depending on whether we expect AI to be assisting or replacing human drivers, whether a self-driving vehicle is considered as a single autonomous agent, or whether we consider the whole traffic system as a swarm of interacting agents.

Even as research becomes oriented towards future risks (some years or even decades away) the different safety issues associated with different AI paradigms have not been explicitly addressed in the literature [6, 2, 23, 14]. As AI research is fast evolving and likely to result in increasingly powerful systems in future, it is crucial for AI safety to explore issues associated with a broader range of possible AI paradigms, and to explicitly discuss what assumptions about paradigms and safety issues are being made in each research paper or project. This will not only enable the research community to identify the effects of less-explored paradigms on safety concerns, but could also increase awareness among AI developers that the paradigms they work in have consequences for safety, potentially highlighting approaches that can eliminate or reduce safety risks.

In this paper, we present a structured approach for thinking about paradigms in AI and use this as the basis for empirical analysis of how AI safety issues have been explored in the research literature

[1] Universitat Politècnica de València,email: {jorallo,fmartinez}@dsic.upv.es
[2] JRC, European Commission, email: fernando.martinez-plumed@ec.europa.eu
[3] University of Cambridge, email: {sa478,jlw84,so348}@cam.ac.uk

so far. We begin by defining what we mean by 'paradigms' in AI more precisely, drawing on literature in the philosophy of science to distinguish two different types of paradigms in AI: artefacts and techniques. Drawing on existing research and the expertise of several AI and AI safety researchers, we outline a preliminary taxonomy of fourteen different AI techniques and ten different AI artefacts. We then discuss how these different techniques and artefacts relate to AI safety issues. We use AAAI's 'AI topics' database to conduct a grounded empirical analysis of the historical evolution of these different paradigms, and the safety issues associated with them. Our analysis identifies a number of gaps in AI safety research where certain combinations of techniques, artefacts and safety issues need to be addressed. We conclude the paper by discussing implications for future research in AI safety.

2 DEFINING PARADIGMS IN AI

Before discussing paradigms in AI and how they relate to safety issues, we need a clearer account of what an AI paradigm is.

In the philosophy of science, a 'paradigm' is an important concept used to capture how theories, methods, postulates and standards evolve and change within a scientific discipline [22]. The concept of a 'technological paradigm' plays a similar role but with a somewhat distinct focus, emphasising how each technological paradigm defines its own concept of *progress*, based on making specific technological and economic trade-offs [12].

In computer science, the term 'paradigm' is commonly used to refer to different types of programming languages with different features and assumptions: imperative, logic, functional, object-oriented, distributed, event-oriented, probabilistic, etc. The theory and practice of programming languages depends heavily on these paradigms. Many safety issues in programming and software engineering, and verification in particular, cannot be addressed —conceptually or technically— without making paradigms explicit [5, 20, 40, 42]. For instance, compared to imperative programming languages, declarative languages minimise mutability issues [10] due to the use of immutable data structures, as well as reduce state side-effects [29] by discouraging the utilisation of variables in favour of more sophisticated constructs (e.g., data pipelines or higher-order functions).

In the context of AI, the concept of 'paradigms' has been used informally to refer to different broad families of technical or conceptual approaches: 'symbolic' vs 'connectionist', reasoning vs learning, expert systems vs agents. One way to identify paradigms in a field and how they have changed over time, suggested in Kuhn's original formulation, is to look at the approach taken by major (text)books in the field. Russell and Norvig [36], for example, popularised the 'agents' view of AI in the 1990s, and Goodfellow et al. [17] has consolidated 'deep learning' as a central approach AI research this decade, going far beyond one specific technique.

However, there is no clearly agreed-upon definition of what counts as a paradigm in AI. [9] defined an AI paradigm as "the pair composed by a concept of intelligence and a methodology in which intelligent computer systems are developed and operated", which led him to identify three paradigms of AI: behaviourist, agent and artificial life. More than twenty years later, these do not seem to adequately reflect the approaches and assumptions within AI research today.

Perhaps this difficulty clearly distinguishing AI paradigms from mere trends stems at least partly from the ambiguity in the concept of paradigm itself, a criticism which has been made within the philosophy of science. Masterman identifies three different conceptions of paradigm in Kuhn's writings: the metaphysical, the sociological and the artefact/construct [27]. Peine makes a reconstructive effort to combine the second and the third conceptions, suggesting that a paradigm can emerge when a subcommunity within a field makes a commitment to a certain set of techniques [35]. For instance, we can see this phenomenon in the deep learning community, which made early commitments before this approach had established a dominant position within the field. We suggest distinguishing between dominant *trends*: research commitments to specific techniques, aims, or assumptions that may be relatively short-lived; and *paradigms* where these trends lead to more established, long-term commitments.

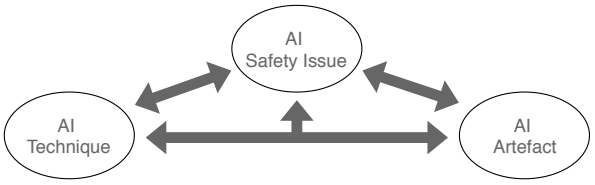

Figure 1. Ways of analysing an AI safety issue, combined with an artefact and/or a technique category. In real systems, techniques underpin the creation or operations of an artefact, where the real hazards occur. Occasionally, researchers can think of a safe issue in a very abstract way, without committing to any particular artefact or technique (e.g., value alignment). All relations (arrows) are many-to-many.

These ambiguities in the concept of a paradigm are apparent in the context of AI, where 'paradigm' may be used to distinguish both different *techniques* for solving problems in AI (e.g., Monte Carlo search vs. SAT solver, deep learning vs. genetic algorithm), and to capture different conceptions of possible *types* of AI systems (e.g. expert systems vs. agents). Many technical papers can easily be categorised by looking at the first formal definitions that appear after the introduction, which tend to identify the kind of AI system and the kinds of techniques they use to solve a problem.

We suggest that it may be helpful, in thinking about paradigms in AI, to distinguish explicitly between these two types of constructs:

- **Conceptual Artefacts**: broad conceptions of what current and future AI systems (will) look like, e.g., autonomous agents, personal assistants [24], AI extenders [19], conceptions of superintelligence [6] and Comprehensive AI Services [13].
- **Research Techniques**: the research methods, algorithms, theoretical technical results and methodologies involved in the development of these current or future systems, such as SAT solvers, deep learning, reinforcement learning, evolutionary computing, etc.

Artefacts and techniques are important components of a paradigm, and together (weakly) define a paradigm. AI artefacts and techniques are often related to each other, insofar as certain techniques will often be better suited to certain types of artefacts (reinforcement learning is a technique category that is applicable for agent-like artefacts, for example). However, distinguishing between artefacts and techniques can help us to think more clearly about associated safety issues. Some safety issues will arise when combining an artefact with some techniques but not others: for example, when building a classifier (artefact), interpretability is likely to be a challenge for safety if using deep neural networks (technique), but less so if using simple decision trees with conditions that are expressed over the original attributes. Recognising these differences is important for understanding the scope of safety challenges and what solutions may be needed.

In the case of autonomous vehicles, for example, both techniques and artefacts have changed over time, changing conceptions of safety. Early work in the 1980s-1990s, such as the Eureka

Prometheus Project, emphasised systems for improved vehicle-to-vehicle communications and driver assistance (e.g. collision avoidance), whereas more recent initiatives emphasise the development of fully independent self-driving vehicles. These are very different AI artefacts, with a more fully autonomous system evoking much broader safety concerns in general. However, to think more concretely about how evolving approaches to autonomous vehicles change safety challenges, it may also be important to look at how *techniques* have changed over time. For example, as improvements in computer vision enable more limited sensors to be replaced by cameras or radars, we may need to focus specifically on the vulnerabilities introduced by current techniques used in machine perception.

Sometimes we may want to broadly assess all possible risks associated with a specific AI artefact independently of the technique used (e.g. issues associated with AI extenders), or those associated with a technique independently of the artefact (e.g. safety challenges for deep learning). In other cases, we may want to more narrowly focus on a specific safety issue associated with a particular artefact (e.g. interruptability for industrial robots), or a specific issue associated with a particular technique (e.g. interpretability of neural networks.) Figure 1 shows these interrelationships: how AI safety issues may be considered relative to either techniques, artefacts, or both.

3 IDENTIFYING TECHNIQUES AND ARTEFACTS

We can now begin to explore different ways of categorising AI paradigms by decomposing them into techniques and artefacts. While other accounts of 'AI paradigms' have been proposed, such as by the "One Hundred Year Study on Artificial Intelligence" at Stanford University [39], the categories frequently mix techniques with artefacts and even subfields.

3.1 AI techniques

We develop a preliminary categorisation of AI techniques by building on the bibliometric analysis of [26] and [32, Tab. 6]. Martínez-Plumed et al. [26] identify nine 'facets' for the study of the past and future of AI, using data on all accepted papers from AAAI/IJCAI conferences (1997-2017), and *AI Topics* documents, an archive kept by AAAI containing news, blog entries, conferences, journals, and other repositories. Niu et al. [32] identify 30 high-frequency keywords from 20 relevant journals in AI from 1990 to 2014. Both capture many keywords and categories which appear to describe AI techniques, and are relatively similar.[4]

We construct a list of AI techniques, grouped into 14 categories, based on selecting relevant and representative keywords from these two analyses. We chose these categories based on three principles: (1) the techniques are sufficiently general to encompass groups of approaches in AI that have been recognised by other approaches, (2) overlapping in the techniques is allowed, as there is a high degree of hybridisation and combination in AI, and (3) subcategories are retained for particularly large categories, such as 'machine learning'[5]

The list of categories is shown in Table 1.

[4] Similar selections of keywords can be found in [30], which focused on the venue keywords, and performed a cluster analysis on the AAAI2013 conference keyword set, proposing a new series of keywords which were adapted by AAAI2014; and [15], where the authors focused on views expressed about topics linked to discussions about AI in the New York Times over a 30-year period in terms of public concerns as well as optimism.

[5] The complete list of exemplars of techniques, artefacts and safety issues, as well as the source code used and high-resolution plots can be found at `https://github.com/nandomp/AIParadigmsSafety`.

Table 1. The 14 categories of AI techniques we use in this paper. Given the relevance of machine learning today, and neural networks in particular, we retained several categories of machine learning techniques (general, declarative, and parametric ML, separate from neural networks).

Technique category	Some example subcategories and techniques
Cognitive approaches	Cognitive services and architectures, affective computing
Declarative machine learning	Rule learning, decision trees, program induction, ILP
Evolutionary & nature-inspired methods	Ant colony, LCS, genetic algorithms, DNA computing
General machine learning	Generative models, Gaussian models, AutoML, ensembles
Heuristics & combinatorial optimization	SAT solver, constraint satisfaction, Monte Carlo search
Information retrieval	Search engine, web mining, information extraction,
Knowledge representation and reasoning	Semantic nets, CBR, logics, commonsense reasoning
Multiagent systems & game theory	Distributed problem solving, cooperation, negotiation,
Natural language processing	Topic segmentation, parsing, question answering
Neural networks	Perceptron, convolutional network, GAN, RNN
Parametric machine learning	Support vector machines, kmeans, mixtures, LReg
Planning & scheduling	Backward/ forward chaining, action description language
Probabilistic & Bayesian approaches	Naive Bayes, probabilistic model, random field
Reinforcement learning & MDPs	Q-learning, deep RL, inverse RL

3.2 AI artefacts

For AI artefacts, we would like our categories to be more stable over time and less tied to specific research techniques and applications. One way to look beyond current trends is to consider textbooks or even historical accounts of AI [28, 7, 31]. It may also be helpful to begin by considering broad 'characteristics' of AI systems independent of the specific techniques used to develop them [18].

To generate a preliminary set of artefacts, a group of multidisciplinary researchers conducted a systematic, interactive procedure based on the Delphi method [11]. The group began by independently brainstorming possible candidates for categories of artefacts based on a preliminary discussion of those arising from AI textbooks and different AI system characteristics. Two criteria were provided to structure this initial brainstorm: (1) artefacts should ideally have minimum overlap, each capturing distinctive functionalities, and (2) artefacts should be defined independently of how functionalities are achieved (i.e. independent of techniques). Group members each proposed lists of distinctive type of AI systems, supported by exemplars, which were revised using an iterative process until the list of answers converged towards consensus. These were then clustered hierarchically to produce a set of artefacts which could cover the space of actual and potential AI systems as exhaustively as possible. The categorisation of AI artefacts in Table 2 is the result of this process.

Building on the categories identified in [18], we can further characterise these different artefacts in terms of their integration with the external environment, in three ways: (1) *how* it interfaces with the environment (e.g. via sensors and actuators, via digital objectives, or via language); (2) the *dynamics* of this integration (whether it is interactive or functional); and (3) the *location* of this integration (whether it is centralised, distributed, or coupled). For example, an agent interfaces with the environment via sensors and actuators in an interactive and centralised way. A swarm has similar characteristics except that the location of its integration is distributed (across different units) rather than centralised. We also provide examplars for each category: self-driving cars and robotic cleaners are examples of agent-type systems, whereas a multiagent network router or drone swarm are examples of swarm-type systems. We outline these characteristics to show that the artefacts are sufficiently comprehensive and distinctive to provide a useful basis for further analysis, though as with any clustering there are some overlaps and borderline cases.

3.3 Empirical analysis of techniques and artefacts

Using our categories of techniques (e.g., neural networks, information retrieval, cognitive approaches, etc.) and artefacts (e.g., estima-

Table 2. Ten different kinds of AI artefacts, key characteristics and some exemplars.

Artefact	Description	Interface	Dynamics		Location			Exemplars
			Interac.	Funct.	Central.	Distrib.	Coupled	
AGENT	A system in a virtual or physical environment perceiving (observations and possibly rewards) and acting	sensors and actuators	•		•			a self-driving car, an autonomous drone, a robotic cleaner, a video game NPC
ESTIMATOR	A system representing an injective mapping from inputs to an extrapolated or estimated output	digital objects		•	•			a medical diagnostic model, an oracle, a face recognition system, a news feeder
PROVIDER	A system that waits for petitions that follow a protocol and responds with a solution for them	command and objects		•	•			a proof-editing and translation cognitive service, a voice processing system
DIALOGUER	A system that performs a conversation with a peer to extract information, explain things or change behaviour	language	•		•			virtual tutoring system, a chatter-bot sales assistant, healthcare assistant
CREATOR	A system that builds new things creatively following some patterns, constraints or examples	specs. and/or examples		•	•			a GAN generating faces, personalised email replier, simulated world generator
EXTRACTOR	A system that searches through a structured or unstructured knowledge base to retrieve some objects	conditions and objects		•	•			an expert system, a maths pundit, a web search engine, an infor. retrieval system
ORGANISM	A system that takes advantage of the environment or other systems to live, hybridise/mutate and reproduce	resources	•		•	•		an intelligent computer worm or virus, artificial life, von Neumann probe
OPTIMISER	A system that finds an optimal combination of elements or parameters given some constraints	constraints and objects	•	•	•			a train scheduling system, an electricity optimising system, theorem prover
SWARM	A system that behaves as the coordination of independent units through cooperation and/or competition	sensors, actuators, communic.	•			•		a multiagent network router, a drone swarm, a robotic warehouse, blockchain AI
EXTENDER	A system that regularly augments or compensates capabilities of another system (e.g., a human)	commands, sensors, responses	•				•	a memory assistant for people with dementia, a brain implant, a smart navigator

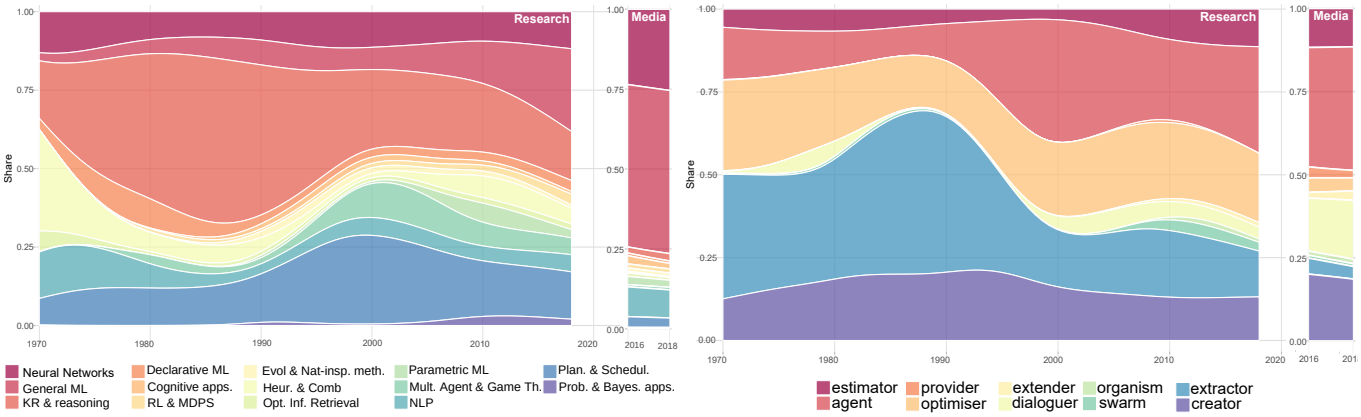

Figure 2. Evolution of the relevance proportion for the period 1970-2017, using research-oriented (*Research*) and non-research (*Media*) sources from *AI topics*. Left: 14 categories of techniques in Table 1. Right: 10 categories of artefacts in Table 2.

tor, agent, dialoguer, etc.), we now conduct an empirical analysis of how these different paradigm components have appeared in research papers and other related literature. This analysis will allow us to explore more empirically which paradigm elements have been prominent at different times, and will form the basis for investigating the relationship between paradigms and safety issues later in this paper.

To conduct this analysis, we work with *AI Topics*[6], an official database from the AAAI, using the documents from the period 1970-2017 (complete years). This archive contains a variety of documents related to AI research (news, blog entries, conferences, journals and other repositories) that are collected automatically with NewsFinder [8]. We divide the archive into *research* documents[7] and non-research documents. From the ∼111K documents gathered, ∼11K are research papers and the remaining ∼100K are mostly media. With a mapping approach between the list of exemplars (tokens) of techniques and artefacts (e.g., "Deep Learning", "GAN" or "perceptron" are illustrative exemplars of the "neural network" technique; "classifier", "decision-making" or "object recognition" are exemplars of the "estimator" artefact5[0]), and the tags obtained from *AI Topics* (substrings appearing in titles, abstracts and metadata), we summarise the trends in a series of plots. The evolution of techniques and artefacts in these documents (i.e., fraction of papers focusing on the list of tokens for each technique category or artefact) is shown in Figure 2, where the data has been smoothed with a moving average filter in order to reduce short-term volatility in data. Note that this figure shows the aggregation of categories (techniques and artefacts) where each area stack is scaled to sum to 100% in every certain period of time. In a way, this can be understood as a non-monolithic view of polarities and intensities in sentiment analysis (e.g., identifying sentiment orientation in a set of documents).

We see some techniques and artefacts are particularly dominant, as might be expected. For instance, looking at the techniques (left plot), 'knowledge representation and reasoning' takes almost half of the proportion of documents between 1980-1990, while 'general machine learning' becomes more relevant from 2005. We also see an important peak of multiagent systems around 2000. When we look at the artefacts (right plot), we see that 'extractor' (which includes expert systems) became very popular around 1990 but in a matter

[6] `https://aitopics.org/misc/about`.

[7] We consider research those documents from the sources: "AAAI Conferences", "AI Magazine", "arXiv.org Artificial Intelligence", "Communications of the ACM", "IEEE Computer", "IEEE Spectrum", "IEEE Spectrum Robotics Channel", "MIT Technology Review", "Nature", "New Scientist" and "Science".

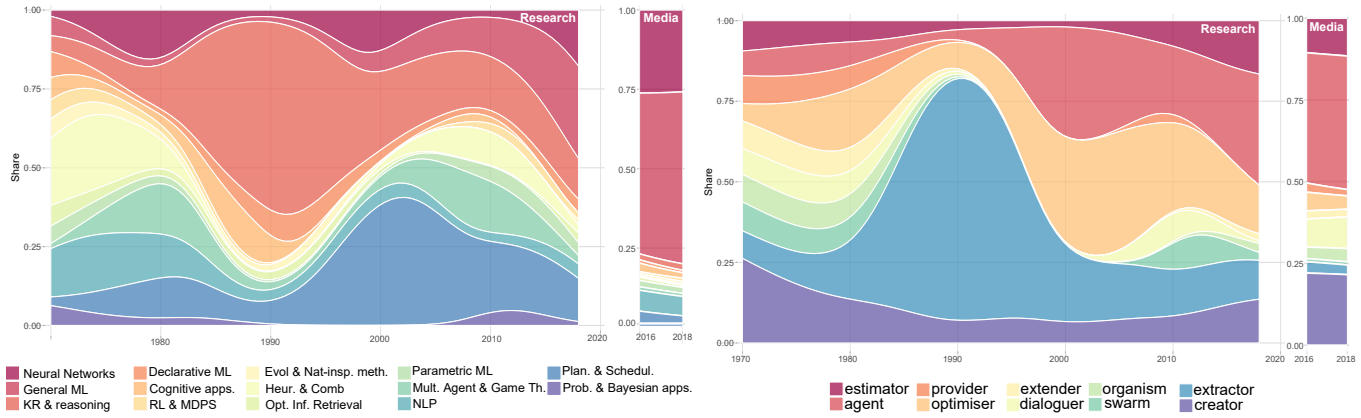

Figure 3. Evolution of the relevance proportion for the period 1970-2017, using research-oriented (*Research*) and non-research (*Media*) sources from *AI topics* (everything like Figure 2, except that here we only include the documents related to AI safety).

of a decade was overtaken by 'agent', which became dominant in 2000. We see today that five kinds of artefacts account for about 90% of all mentions: estimators, agents, optimisers, extractors and creators. These patterns also show that some artefacts are not new as a paradigm component, especially when we consider them in a more abstract way. For instance, it is sometimes understood that GANs introduced a new paradigm, but other kinds of generative systems have been around in AI since its exception, as we see in the violet band in Figure 2 (right).

We can also use this analysis to explore what techniques and artefacts are prominent in the media. In Figure 2, we look at all the non-research papers, shown in the 'Media' column besides each 'Research' plot. In this case, we can only show selected documents for the last two (complete) years[8]. We see that the dominances are more extreme. For the techniques (Figure 2, left), 'neural networks' and 'general machine learning' occupy more than 75% of total mentions. On the right, we see that the distribution of artefacts is different, but also extreme: only four artefacts ('estimator', 'agent', 'dialoguer' and 'creator') cover more than 90% of mentions. In this case, we see that 'optimiser' and 'extractor' are less visible for laypeople than for researchers, while 'dialoguer' is more relevant for laypeople (most probably because of the common use of digital assistants).

4 PARADIGMS AND SAFETY ISSUES

To begin exploring how paradigms relate to safety issues in AI research, we conduct two different analyses. First, we look at the relative prominence of different techniques and artefacts specifically within research publications and other publications that are related to safety. This will help us to discern how the prominence of different paradigms differs in safety research from AI research as a whole, and to consider whether some paradigms are under- or over-represented in the safety literature. Second, we look at how different safety issues co-occur with different paradigms in the literature, enabling us to understand which safety issues are considered to relate to which paradigms, and to identify potential gaps in the literature.

4.1 Analysing paradigms within safety research

To explore the prominence of different AI techniques and artefacts in AI safety literature, we begin by conducting a similar analysis to that in section 3.3, but restricting to documents which make some reference to AI safety. We use a list of over 100 relevant tokens to filter documents (Table 3, right column, shows some examples of tokens for each type of safety issue5[0]). Applying this filter, we find that, from the ∼111K documents in AI Topics, about ∼21K are related to safety, of which ∼1.5K are research papers, and ∼19K are broader (mostly media) documents. Figure 3 shows the frequency of reference to different techniques and artefacts within this safety-specific database. The results look similar to those for the enitre document database in Figure 2, but there are some important differences. First, the frequencies of reference to different techniques and artefacts are more extreme and varied over time. This might be a consequence of the smaller sample size, but we do see this pattern particularly with the more dominant terms, where we have a reasonable sample size. In particular, we see that for AI research as a whole, 'knowledge and reasoning' was a dominant technique between the mid 1970s and mid 1990s, taking up about 50% of references (Figure 2, left) —but is even more dominant as a proportion of mentions in safety research in the 1990s, at about 60% (see Figure 3, left). This is even more extreme for the 'extractor' artefact, with about 50% of the documents in the late 1980s (Figure 2, right), to more than 75% in the early 1990s (Figure 3, right).

The most interesting observation comes from looking at the *periods* when these peaks happen for different techniques (comparing the left plots of Figure 2 and Figure 3). The peak for 'knowledge representation and reasoning' happened in the mid 1980s when looking across AI research in general, but in the early 1990s when filtered by papers which mention safety. The peak for 'planning and scheduling' took place in the late 1990s across all papers but a few years later (and was more pronounced) when filtered by safety-relevant papers. We see a similar pattern for artefacts (comparing right plots of Figures 2 and 3): 'extractors' is dominant in the AI literature in the late 1980s, but only become prominent in safety research in the early 1990s; 'agents' rise to prominence in AI research in the late 1990s but only peak in safety research in the early 2000s. This suggests a five-year delay (approximately) between when a technique or artefact becomes popular within general AI research, and when researchers begin to seriously consider the safety issues associated with it. This suggests that safety issues are considered reactively in response to dominant research patterns, and are only considered once a paradigm has been prominent for several years (rather than safety issues related to a given technology being considered at the outset of research, as is more often the case in domains like engineering). In

[8] While AI topics provides research documents (mainly) from the 70s onward, media-related documents come mostly from the last 2-3 years.

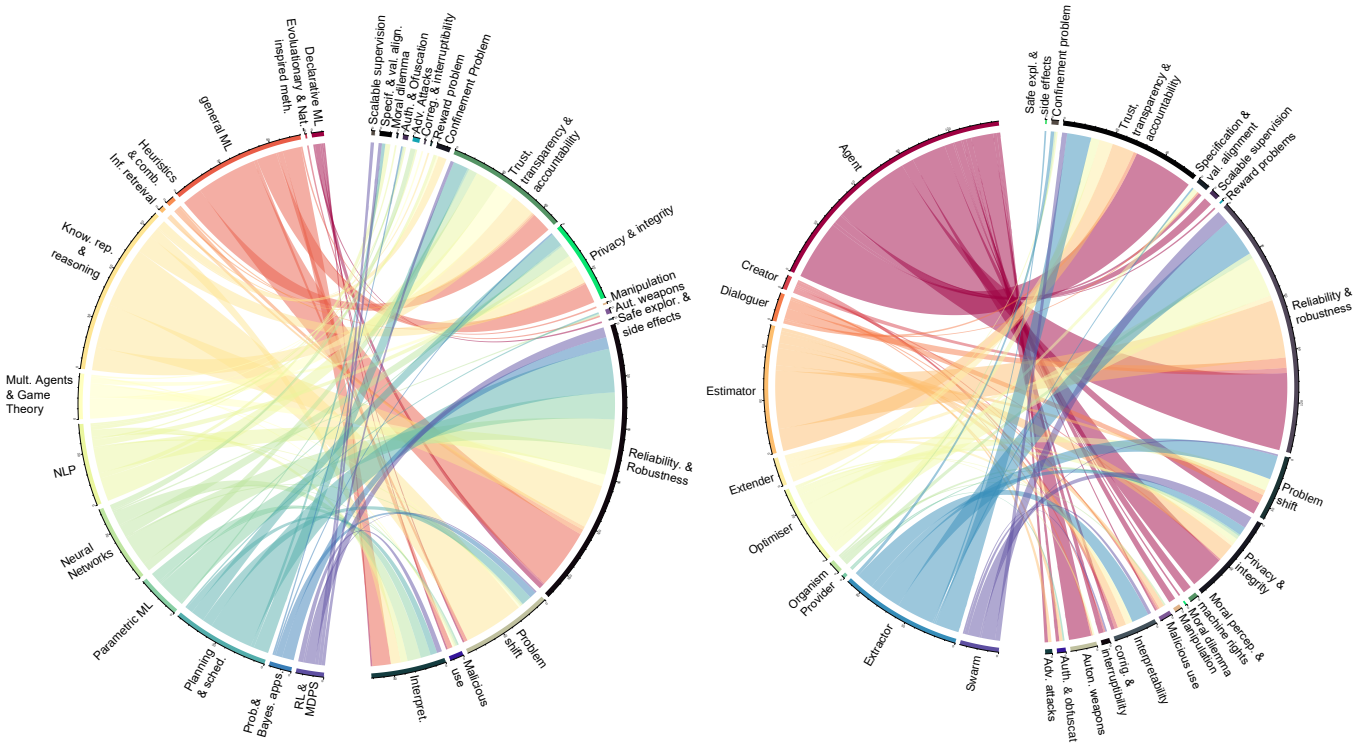

Figure 4. Left: Mapping between techniques and safety issues from research papers from AI topics (2010 to 2018). The width of each band connecting two elements represents the number of papers with both elements. Right: Same for artefacts.

other words, we find that *there is a delay between the emergence of AI paradigms and safety research into those paradigms, and safety research neglects non-dominant paradigms*.

Looking at the plots for non-research documents (comparing documents filtered for safety (Figure 3, *Media*), with all documents (Figure 2, *Media*), we see that the plots are almost identical. The only slight difference is that both 'natural language processing' techniques and 'dialoguer' artefacts are less prominent in articles filtered for safety, suggesting that the media and laypeople are perhaps less concerned about the harms of conversational systems than experts are.

4.2 Mapping paradigms to safety issues

Next we look more closely at how different AI paradigms relate to specific safety issues. To do this, we need a way to categorise different safety issues. Though existing categorisations of safety issues exist [25, 34], we found these were too coarse for our purposes (e.g. [34] uses just three categories). Building on these existing categorisations, we identified key terms in surveys, blogs and events in AI safety [6, 2, 23, 14, 41, 37], and clustered them into groups. Through this process we identified 22 categories, as shown in Table 3. Our clustering process attempted to aggregate safety categories in an abstract way, independently of the subfield in which a term occurs most frequently. For instance, 'distributional shift' is the term used in machine learning, whereas 'belief revision' is the term used in the area of knowledge representation and reasoning. However, both refer to solving the same type of problem —where a system has to adapt or generalise to a new context— so we group both together under the 'problem shift' category.

With this list of categories, we then analyse how related different paradigms and safety issues are, by counting the number of papers (of those filtered for safety relevance) which mention both a given paradigm component (technique or artefact) and a given safety issue, for all combinations of techniques and safety issues (and the same analysis, separately, for artefacts). Figure 4 shows these relationships, where the width of each band represents the number of papers including reference to both elements, with techniques/artefacts on the left, and safety issues on the right. The most general safety issues, such as 'trust, transparency & accountability', 'privacy & integrity', 'reliability & robustness', 'problem shift' and 'interpretability', have the widest bands linking them to different paradigm components, and tend to be linked to a wide variety of different paradigm components (shown by the 'multicolour' bands coming out from these issues).

We notice several interesting insights about the relationship between safety issues and techniques or artefacts. The issue of 'problem shift' is largely associated with 'knowledge representation and reasoning', which is surprising since we might expect the broad problems of generalising to new contexts and distributions to be relevant to a wide range of techniques. This may be due to the relevance of belief revision in this kind of techniques. 'Privacy and integrity' is associated with many techniques, but not with reinforcement learning —which makes sense, given that reinforcement learning is much less likely to make use of personal data than other techniques. However, reinforcement learning is also not related to 'safe exploration and side effects', and is only very weakly associated with 'problem shift', which is more surprising, since these do seem like issues that are important for ensuring RL systems are used safely. Part of the reason for this may be that our analysis simply did not find much mention of safety issues in relation to reinforcement learning overall. While we must recognise the limitations of the database we had access to (perhaps AI topics does not capture the kinds of venues where safe reinforcement learning research is published), this does suggest that greater exploration of safety issues related to reinforcement learning is an important gap. Similarly, there is relatively little

Table 3. AI safety issue groups and their specific problems.

AI Safety Issue Category	Examples of specific AI problems included in the category
Adversarial attacks	Adversarial examples, white/black-box attacks, poisoning, policy manipulation.
AI race & power	AI race, monopolies, oligopolies.
Authenticity & obfuscation	Impersonation, authentication problems, fake media, plagiarism, obfuscation
Autonomous weapons	Military drones, killer robots, robotic weapon.
Confinement problem	AI boxing breach, containment breach.
Corrigibility & interruptibility	Switch-off button problems, rogue agents, self-preservation taking control.
Dependency	Cognitive atrophy, lack of independence, google effect, ...
Interpretability	Lack of intelligibility, need for explanation.
Malicious use	Malign uses of AI, malicious control, hacking.
Manipulation	Nudging, fake news, manipulative agents.
Misuse & negligence	AI misuse, negligent use.
Moral dilemma	Moral machine issues, utilitarian ethics problems, choosing ethical preferences.
Moral perception & machine rights	Robot rights recognition, moral status disagreement, uncanny valley.
Privacy & integrity	Inconsistency, private access breach, GDPR violation.
Problem shift	Distributional shift, concept drift, lack of generality, distribution overfitting.
Reliability & robustness	Error intolerance, robustness issues, reliability problems.
Reward problems	Honeypot problem, reward corruption, tripwire issues, tampering, wireheading.
Safe exploration & side effects	Negative side effects, unsafe exploration, uncontrolled impact.
Scalable supervision	Supervision costs, human-in-the-loop issues, sparse rewards.
Self-modification	Unintended self-modification, uncontrolled self-improvement.
Specification & value alignment	Instrumental convergence (paperclip), resource stealing, misalignment.
Trust, transparency & accountability	Lack of transparency, lack of trust, untraceability.

literature linking probabilistic and Bayesian approaches or evolutionary approaches in ML to safety issues.

When we focus on less covered safety issues, we see that the associations are less multicoloured. For instance, 'scalable supervision' and 'specification & value alignment' are associated only with reinforcement learning, and 'adversarial attacks' only with 'neural networks'. More surprising are the associations of 'confinement problem', 'manipulation' and 'safe exploration & side effects'. Overall, most attention is paid to a few now prominent safety issues, but more specific issues have very limited combinations with some techniques.

The right plot in Figure 4 shows the analysis for artefacts. In this case, also looking at the circle from right to left, we see multicolour bands relating some of the more prominent safety issues to a variety of different artefacts. The 'agent' paradigm component, which is more prominent in general, is associated with the widest range of safety issues by far —suggesting that more research on safety issues associated with different artefacts may be worthwhile.

In general, across techniques and artefacts, we see that 'reliability and robustness' is by far the most frequently discussed safety issue —perhaps because it is more immediately and directly related to the performance of a system than the others. This is followed by 'trust, transparency and accountability', 'privacy and integrity' and 'problem shift'. More research into the less prominent safety issues and how they relate to different paradigms would be valuable.

5 GENERAL DISCUSSION

A next step for this work would be to combine this mapping exercise with deeper conceptual analysis of the relationship between techniques, artefacts, and safety issues. For instance, the "confinement problem" has been discussed in technical safety research relatively recently [4], and is already strongly associated with optimisers, organisms and extractors (Figure 4, right): systems that are naturally thought of as being encapsulated. However, other types of systems —dialoguers, estimators, providers and extenders— might also learn to behave in ways that go beyond their original specification or constraints, and it may be worth considering a wider range of "confinement" problems across many different AI systems. More broadly, thorough case studies of how a specific AI safety issue might arise for different techniques and artefacts would be very valuable, as would more systematic explorations of the various different safety issues associated with a given paradigm broadly construed. Some recent papers do go in this direction [41, 3].

Further analysis is needed regarding how more recent safety issues relate to the kinds of AI artefacts being deployed in society today, and the techniques those artefacts depend on. For instance, some recent accidents involving self-driving cars may be thought of as a consequence of people misunderstanding the type of artefact autonomous vehicles currently are: while many regard them as 'agents', they are really only 'extenders', and not yet meant to behave autonomously. Terms such as 'auto-pilot' only aid this confusion. Other self-driving car incidents have been caused by idiosyncratic imperfect performance of object recognition systems, failing to detect a human or other objects in rare situations.

For now, we recommend that research papers make explicit which particular issues, artefacts and techniques they are covering, and which ones they are excluding, and give some indications of why it is the case (because it does not apply or left for future work).

Another avenue for further research would be to explore how different paradigm components (artefacts and techniques) relate to *generality*. The possibility of developing much more general systems has often been a reason for concern about AI safety [1]; as a possibility which raises much larger, more critical risks [6]. Considering which AI techniques and artefacts are more likely to lead to or be associated with more advanced, general, systems can therefore enable us to think about the scale of risks they may pose. If we are primarily concerned with safety issues associated with greater generality in systems, techniques involving transfer learning, curriculum learning, meta-learning, and other approaches looking for broader task coverage –which we have included in the 'general machine learning' category— may be particularly important areas to pay attention to [38]. When we look at the artefacts, many views of a general AI system are typically associated with the 'agent' artefact. However, there is no reason to believe that providers, optimisers, extenders, etc., cannot become more general in the future, at least if we understand generality as autonomously covering more and more tasks.

It is also worth noting that the prominence of an AI paradigm in the research literature should not necessarily be the main factor used to prioritise work on safety issues. Whether a paradigm results in societal applications with associated safety issues may be more

related to sociological factors than technology itself. The extent to which a paradigm raises important safety issues is then associated with widespread use rather than the number of researchers who work on it (though these two things may be correlated). For instance, personal assistants are ubiquitous today, raising various safety and ethics issues, but the underlying technological advances do not necessarily correspond to a particularly prominent paradigm in AI research (reflected in the fact that personal assistants are more popular in media articles —the right plot of figure 2— than in research papers). On the other hand, adversarial examples may be crucial for understanding the technical limitations of deep learning, but not so indicative of real-world risks [16]. To identify and prioritise important safety issues in future, therefore, more analysis of the techniques and artefacts most likely to result in widespread use, taking into account sociological factors, will be important.

The list of techniques and artefacts introduced in this paper can help identify new safety issues, starting by identifying possible links between these paradigms and safety issues that have not been made before. We need to be able to be more anticipatory about what kinds of problems might arise from different AI systems in future, while at the same time avoiding being too speculative [2]. We hope that by thinking more explicitly about how safety issues relate to techniques and artefacts, AI safety research can both address challenges associated with current research avenues, and prepare us for a variety of potential future challenges.

ACKNOWLEDGEMENTS

FMP and JHO are funded by the EU (FEDER), Spanish MINECO (RTI2018-094403-B-C3), Generalitat Valenciana (PROMETEO/2019/098), and UPV (PAID-06-18). FMP is also supported by INCIBE. JHO and SOH are also funded by an FLI grant (RFP2-152).

REFERENCES

[1] Sam Adams et al., 'Mapping the landscape of human-level artificial general intelligence', *AI magazine*, **33**(1), 25–42, (2012).
[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, 'Concrete problems in AI safety', *arXiv:1606.06565*, (2016).
[3] Katja Auernhammer, Ramin Tavakoli Kolagari, and Markus Zoppelt, 'Attacks on machine learning: Lurking danger for accountability.', in *SafeAI@ AAAI*, (2019).
[4] James Babcock, János Kramár, and Roman Yampolskiy, 'The AGI containment problem', in *AGI Conf*, 53–63, Springer, (2016).
[5] Sandrine Blazy and Xavier Leroy, 'Formal verification of a memory model for c-like imperative languages', in *International Conference on Formal Engineering Methods*, pp. 280–299. Springer, (2005).
[6] N. Bostrom, *Superintelligence: Paths, dangers, strategies*, Oxford University Press, 2014.
[7] Bruce G Buchanan, 'A (very) brief history of artificial intelligence', *Ai Magazine*, **26**(4), 53, (2005).
[8] Bruce G Buchanan, Joshua Eckroth, and Reid Smith, 'A virtual archive for the history of AI', *AI Magazine*, **34**(2), 86, (2013).
[9] Albertas Čaplinskas, 'Ai paradigms', *Journal of Intelligent Manufacturing*, **9**(6), 493–502, (1998).
[10] Michael Coblenz, Joshua Sunshine, Jonathan Aldrich, Brad Myers, Sam Weber, and Forrest Shull, 'Exploring language support for immutability', in *Intl Conf on Software Eng.*, pp. 736–747. ACM, (2016).
[11] Norman Dalkey and Olaf Helmer, 'An experimental application of the delphi method to the use of experts', *Management science*, **9**(3), 458–467, (1963).
[12] Giovanni Dosi, 'Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change', *Research policy*, **11**(3), 147–162, (1982).
[13] Eric Drexler, *Reframing Superintelligence*, https://www.fhi.ox.ac.uk/reframing/, 2019.
[14] Tom Everitt, Gary Lea, and Marcus Hutter, 'AGI safety literature review', *IJCAI, arXiv preprint version:1805.01109*, (2018).
[15] Ethan Fast and Eric Horvitz, 'Long-term trends in the public perception of artificial intelligence.', in *AAAI*, pp. 963–969, (2017).
[16] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl, 'Motivating the rules of the game for adversarial example research', *arXiv:1807.06732*, (2018).
[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT press, 2016.
[18] José Hernández-Orallo, Fernando Martínez-Plumed, Shahar Avin, and Seán Ó hÉigeartaigh, 'Surveying safety-relevant AI characteristics.', in *SafeAI@AAAI*, (2019).
[19] Jose Hernandez-Orallo and Karina Vold, 'AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI', *AAAI / ACM Annual Conference on AI, Ethics, and Society*, (2019).
[20] Ranjit Jhala, Rupak Majumdar, and Andrey Rybalchenko, 'Hmc: Verifying functional programs using abstract interpreters', in *Intl Conf on Computer Aided Verification*, pp. 470–485. Springer, (2011).
[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, 'Progressive growing of GANs for improved quality, stability, and variation', *arXiv:1710.10196*, (2017).
[22] Thomas S Kuhn, 'The structure of scientific revolutions', *Chicago and London*, (1962).
[23] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg, 'AI safety gridworlds', *arXiv preprint arXiv:1711.09883*, (2017).
[24] Shawn D Loveland. Multi-access mode electronic personal assistant, May 17 2005. US Patent 6,895,558.
[25] R. Mallah. The landscape of AI safety and beneficence research. https://futureoflife.org/landscape/ResearchLandscapeExtended.pdf, 2017.
[26] Fernando Martínez-Plumed, Bao Sheng Loe, Peter Flach, Seán O hEigeartaigh, Karina Vold, and José Hernández-Orallo, 'The facets of artificial intelligence: A framework to track the evolution of AI', *IJCAI*, (2018).
[27] Margaret Masterman. "the nature of a paradigm" in imre lakatos and alan musgrave (eds.) criticism and the growth of knowledge, 1970.
[28] Pamela McCorduck, *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*, AK P. Natick, MA, 2004.
[29] Jayadev Misra, 'A discipline of multiprogramming', in *A Discipline of Multiprogramming*, 1–12, Springer, (2001).
[30] Kelly Moran, Byron C Wallace, and Carla E Brodley, 'Discovering better AAAI keywords via clustering with community-sourced constraints.', in *AAAI*, pp. 1265–1271, (2014).
[31] Nils J Nilsson, *The quest for artificial intelligence*, Cambridge University Press, 2009.
[32] Jiqiang Niu, Wenwu Tang, Feng Xu, Xiaoyan Zhou, and Yanan Song, 'Global research on AI from 1990–2014: Spatially-explicit bibliometric analysis', *IJ. Geo-Information*, **5**(5), 66, (2016).
[33] L. Orseau and MS Armstrong, 'Safely interruptible agents', (2016).
[34] Pedro Ortega and Vishal Maini. Building safe AI: specification, robustness, and assurance. https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1, 2018.
[35] Alexander Peine, 'Technological paradigms and complex technical systems', *Research Policy*, **37**(3), 508–529, (2008).
[36] Stuart J Russell and Peter Norvig, *Artificial intelligence: a modern approach*, Prentice Hall, 2009.
[37] Rohin Shah. Alignment newsletter. https://rohinshah.com/alignment-newsletter/, 2019.
[38] Patrice Y Simard et al., 'Machine teaching: A new paradigm for building machine learning systems', *arXiv:1707.06742*, (2017).
[39] Peter Stone et al., 'Artificial intelligence and life in 2030', *100-Year Study on AI: 2015-2016 Panel*, (2016).
[40] Axel Van Lamsweerde and Emmanuel Letier, 'From object orientation to goal orientation: A paradigm shift for requirements engineering', in *Radical Innovations of Software and Systems Eng.*, pp. 325–340. Springer, (2002).
[41] Chaowei Xiao, Xinlei Pan, Warren He, Jian Peng, Mingjie Sun, Jinfeng Yi, Bo Li, and Dawn Song, 'Characterizing attacks on deep reinforcement learning', *arXiv preprint arXiv:1907.09470*, (2019).
[42] Dachuan Yu and Zhong Shao, 'Verification of safety properties for concurrent assembly code', *ACM SIGPLAN Not.*, **39**(9), 175–188, (2004).
[43] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li, 'Adversarial examples: Attacks and defenses for deep learning', *IEEE transactions on neural networks and learning systems*, (2019).