Incremental Autonomy

A Vision for Next-Generation AI

M Rovatsos, May 2022

Introduction

Despite enormous advances in AI, the vast majority of systems available today are still limited to performing relatively simple functions, and those that exhibit human-level performance at more complex tasks have been largely limited to simulated settings. At the same time, recent progress has been accompanied by increasing concerns around the ethical risks posed by AI and its environmental and societal sustainability, which are exacerbated by the use of opaque systems developed using very large amounts of data and computation.

We posit that the underlying cause for these problems is an exaggerated focus on *monolithic* AI systems, rather than incremental approaches where more complex systems can be built systematically by reusing, adapting, and integrating existing components. This focus often goes hand in hand with an excessive emphasis on training systems to achieve (super)human performance on a single task, which is bound to create opaque systems as increasing system performance takes priority over other design objectives such as interpretability or reusability. If we follow the currently dominant paradigm, it is likely that more and more data and compute power will be needed as we try to solve increasingly complex problems, to the point where we may come up against hard limits in terms of the kinds of problems that can be solved by a single system. Moreover, this reliance on data and compute is environmentally unsustainable, and training a whole new system from scratch for every possible task can be particularly wasteful when tasks are similar. We can also expect that this paradigm will exacerbate privacy, power imbalance, and other ethical issues.

We present an alternative vision for engineering next-generation AI systems that aims to overcome these problems, and which is based on enabling increasing levels of autonomy in AI systems by integrating individual component capabilities through an incremental process controlled by iterations of designing, composing, and validating assemblages of existing components. We argue that this approach has the potential to unlock novel AI capabilities, and provides a sustainable, responsible, and productive pathway to developing next-generation AI systems.

Why Autonomy?

While many dispute that a future where AI systems act autonomously is desirable, we believe that shifting the focus of AI development towards autonomy has the potential to address many of the limitations and risks associated with current AI technologies. From the standpoint of application benefit, if we want AI systems to perform increasingly complex tasks on our behalf, they will arguably have to exhibit and control a range of flexible behaviours independently, respond to events in their environment, and take action without constant human intervention. This, in turn, will necessitate degrees of complexity and encapsulation of functionality that will make it impossible for human users to examine and understand all internal details of a system. In fact, AI systems available today already make this practically impossible in many cases, even if their functionality is relatively simple.

Given this, the real question becomes one of how we can achieve effective control of these systems in the face of increasing autonomy. We argue that a productive relationship between users and AI systems does not require maintaining a full understanding of how these systems work, and that enforcing this requirement would even limit the future development of new AI capabilities. Instead, we need to establish engineering methodologies aimed at increasing the autonomy of a system incrementally while rigorously validating the ability of human users to control its behaviour effectively.

Our vision for developing such a methodology is based on the notion of *incremental* autonomy, which (1) views autonomy as a variable property of AI systems along a spectrum from full human control to fully independent action (and which applies to both physical and software AI-driven systems), (2) focuses on the systematic *integration* of component systems to compose more complex behaviours out of simpler ones, and (3) views meaningful human-AI *interaction* as the key criterion that should guide the development of increasingly autonomous systems.

It is important to acknowledge that almost no currently available systems provide sufficient assurances to warrant full autonomy, and that, in many cases, we may well decide that full autonomy is not possible or desirable at all. However, we believe that iteratively exploring increasing levels of autonomy provides a safe

and responsible pathway to assessing and mitigating the risks and feasibility of embedding future AI systems in society. It is also worth emphasising that we do not view our approach as a pathway to so-called "artificial general intelligence", or as part of debates around ascribing rights, responsibility, consciousness or free will to AI systems. We take a pragmatic approach that asks how we can create useful capabilities in tools that may require replicating elements of human intelligence, and aims to exploit the potential benefits of advanced autonomy in safe and responsible ways.

A New Focus for Design

The starting point for developing a design methodology for incremental autonomy is to ask how we can build systems that exhibit the levels of flexibility of behaviour that would warrant delegating increasingly complex tasks to them. If we want to move away from a "monolithic" approach that involves developing custom architectures and algorithms fine-tuned to achieve the desired performance on a very specific task, a natural approach is to combine and reuse existing components. However, at present, most of these AI components are typically not developed with integration in mind, and achieving a degree of *compositionality* when putting them together is a concern that has hitherto received little attention.

To illustrate the complexities involved in composing AI systems, consider the example of a legged robot that can safely move around complex physical spaces, and which we wish to extend by robotic arms to grasp and manipulate delicate objects. The more complex task of carrying such objects while moving around is not just a matter of adding these arms to our robot. The robot may have to reduce its walking speed or increase its minimum distance to obstacles to reduce risks of damaging the object it carries. As another example, assume we wanted to combine a system that monitors rooms in a care home to detect situations where a client might require assistance with a conversational agent on the client's mobile phone used to control assistive devices in the room. This integration could give rise to new failure modes, for example if a misinterpreted user command causes an assistive device to malfunction, and triggers corrective action taken by the user, which is, in turn, wrongly picked up by the room monitoring system as an incident that requires attention.

In focusing on what is *between,* rather than *inside* each component to enable the composition of such complex behaviours out of simpler ones, we can take inspiration from the ways in which we teach people advanced skills once they have mastered simpler ones. Typically, this involves making new constraints and dependencies explicit that emerge from the interaction between different individual activities, defining new objectives, metrics, and risks that are relevant to their combination, and careful experimentation to practise and assess the new skill. If successful, this process will allow us to put our confidence in the tutee's ability to perform the new task safely and reliably.

Mapping this onto computational AI systems suggests that design methodologies based on this idea will have to borrow heavily from those developed in other areas of computing such as software and systems engineering to enable iterative and modular software design. Techniques such as aspect-oriented, component- and contract-driven development, integration testing, as well as architectures that involve middleware components and support interoperability will need to be explored and translated to the context of AI systems.

However, we expect new challenges to arise in the application of these techniques in an AI context. As these systems adapt and evolve over time and we move to higher levels of encapsulation of functionality, we expect the role of the developer to shift more towards that of data provider, instructor, and validator. Their role will also become increasingly blurred with that of the end user, who will continue to adapt and influence system behaviour post-deployment, and users will increasingly (individually and/or collectively, in cases where system adaptation is influenced by whole user communities) adopt a role of "prosumers". This raises fundamental issues in terms of anticipating the future behaviour of systems at design time, and new design challenges, for example with regard to the appropriate interfaces needed for those working with autonomous systems to be able to interact appropriately with increasingly complex, evolving behaviours.

Our efforts to develop a conceptual model for an iterative design methodology, which focuses on integrating existing AI components and is controlled through appropriate interaction with human stakeholders, is inspired by the framework of *meta-reasoning* introduced for the design of rational agents in the mid-1990s. Simply put, meta-reasoning provides methods for controlling a computational process (e.g. reasoning, planning, learning) with additional constraints that introduce explicit meta-level control loops. It can be used, for example, to decide whether additional computation is likely to yield significantly better solutions, whether a goal that cannot be attained should be reconsidered, or whether the solvability of a problem should be reassessed. In epistemic terms, it allows additional knowledge about the problem domain to be introduced

"atop" an existing component to control it, given expectations towards and observations made about this component.

We believe that applying this framework to develop new methodologies for the incremental design of autonomous systems has great potential. It mandates an explicit articulation of objectives and constraints we want to impose when integrating components, and putting the machinery in place that will provide effective meta-level control across these components. This will also have implications for the requirements placed on the constituent components themselves, which will need to come with their own explicit definitions of control parameters, behavioural properties, context conditions, and integrity constraints.

If our hypothesis is correct that satisfying these requirements will enable the systematic application of iterative integration methodologies to AI components, a possible implication could be that human-intelligible and -controllable meta-level control is key to unlocking a productive way towards enabling incremental autonomy.

Leveraging Existing Approaches

While our focus on iteration, integration, and interaction aims to provide a new focus for the development of future AI systems through the lens of incremental autonomy, many of its underpinning elements and objectives have been extensively studied across a range of sub-fields of AI for many years, and are in no way novel.

The field of *neuro-symbolic AI*, for example, focuses heavily on integrating different AI methods by extending data-driven statistical approaches with symbolic reasoning components and vice versa. Some of this work follows the tradition of *cognitive architectures*, an area that has proposed blueprints for general-purpose AI architectures that take inspiration from our understanding of the structure of human cognition. The degree to which these lines of work have attempted to integrate existing implemented components varies, but, contrary to our approach, more often than not, they assume that the designer has control over the design of all individual components in the system.

The area of *autonomous agents and multi-agent systems*, on the other hand, has developed a rich arsenal of integration and coordination architectures and algorithms that aim to address many of the issues we are concerned with. These range from methods to integrate individual behaviours in a single AI agent to coordination mechanisms for multiple agents that have no direct control over each other, and whose individual design objectives might even stand in conflict with each other. Methodologically, this area has, however, typically made an assumption that agents are already autonomous, which has limited them either to domains where this can already be realistically assumed, or led them to focus on solving problems for future autonomous systems. To date, few of these techniques have been applied to the kinds of state-of-the-art AI systems in use today.

Within machine learning, similar ideas have been explored in areas such as *federated learning, meta-learning, few-shot learning, and transfer learning*, which generally aim at combining and reusing existing models and/or adapting them with little additional data. These approaches have the potential to provide many of the capabilities we are envisioning. Yet, by and large, they focus on improving the performance and robustness or reducing the amount of data required for training models, rather than the integration of different skills to enable increasing levels of autonomy. Interestingly, many examples of machine learning-based systems that exhibit advanced levels of autonomy (e.g. in game-playing AI) typically integrate state-of-the art machine learning algorithms with other AI components (e.g. stochastic tree search) to obtain the desired capabilities. We view such application-specific (arguably, neuro-symbolic) integration efforts as further evidence for the importance of developing more systematic, generalisable methods to enable autonomy through integration.

An area where application-specific integration has naturally played an even more prominent role in the engineering process is that of *robotics* and other embodied AI systems. Alongside some areas of natural language processing – where autonomy is achieved on tasks such as translation, question answering, or dialogue – robotics is arguably the only area where the "most" autonomous AI systems have been deployed in real-world contexts, and where integration of individual components has been demonstrated successfully. While there have been long-standing efforts to learn more general lessons from these efforts in the field, it is probably fair to say that no general methodologies have been derived that could also be applied to AI systems outside robotics.

Finally, research on *semantic technologies* has developed a range of practical methods to make properties of data, components, and services explicit for the purpose of interconnecting them and providing support for

interoperability and integration. Many of these build on a rich body of work in symbolic AI and database technology, areas which have traditionally focused on compositional representations of information and knowledge. These techniques often also provide much higher levels of human intelligibility than statistical and other so-called "sub-symbolic" methods. However, to date, their use in efforts to enhance autonomous decision making and action in AI systems has been very limited.

We believe that our vision for incremental autonomy will open up significant opportunities for cross-fertilisation and collaboration between these areas in a joint effort to develop new ways of engineering autonomous systems, to which they have valuable methods and insight to contribute. It is important to emphasise that we do not advocate merging or replacing these fields by a new school of thought. Rather, we expect that targeted collaborative efforts that focus on incremental integration of existing AI systems will provide a focus for lateral thinking across communities; and that these new initiatives will benefit from future advances in each of these fields as these continue to emerge.

Towards a Research Roadmap

Work on the proposed vision is still in its very early stages, and further discussion across various AI communities will be needed to translate this high-level vision to a concrete research roadmap. Nonetheless, we can propose a number of general directions that we believe will form the core of a broad, long-term programme of research:

Research that will develop *new ways of assembling AI systems* will be necessary to help overcome the current focus on narrow tasks and enable on-the-fly combination of individual components in modular and expandable systems. On the one hand, this will require a focus on developing *reusable AI* components that work across domains with limited re-training and/or re-configuration, such as transfer learning, autoML/autoRL, or multi-objective learning approaches, as well as research that will enable these systems to provide the descriptions of their functionality that are necessary for reuse and integration. On the other hand, we will have to develop control architectures for integrated AI components that will ideally enable run-time integration, but at least effectively support engineers in their integration efforts. We expect such research to focus on providing component "wrappers", middleware, communication/interaction protocols, distributed architectures, and techniques for defining and validating complex workflows of operation across multiple components.

We also expect research on incremental autonomy to focus on *new ways of adapting systems*, for example through targeted human instruction and demonstration, but also to adapt to the needs of diverse user populations, using, for example, novel approaches to lifelong learning, task generalisation and abstraction. Beyond new methods for data-driven learning and adaptation, we anticipate that this will give rise to the development of new programming models such as high-level languages for controlling and combining AI components, programming by demonstration, new models for customising and configuring systems, and, user-side (including collective, potentially crowdsourced) adaptation.

As increasingly complex AI systems will likely be programmed and configured by developers and users at runtime in more natural and incremental ways, we expect that *new ways of interacting with systems* will emerge as a further research focus. Here, techniques that have been explored in the areas of adjustable autonomy, mixed-initiative AI, and in human-computer interaction more broadly can be deployed to enable on-the-fly personalisation, the development of interfaces to control systems at a behavioural level, and to advance the interpretability of systems with a focus on making increasingly autonomous systems usable, responsive, and safe. Validation and verification of systems to analyse and profile the behaviours of components and integrated systems and provide performance, robustness, and safety guarantees is an important part of managing the uncertainty and risks inherent to the integration of existing AI components, and will likely create new challenges in terms of multi-level and distributed modelling and validation.

Finally, progress in enabling increased levels of autonomy will necessitate the development of *new ways of understanding and using autonomy*, given that whatever levels of autonomy we aim for in a particular context of use, future AI will have to complement human activity meaningfully and safely. We expect that, while this will raise new questions around the ethical and societal risks of autonomous systems, the incremental approach to autonomy has the potential to stimulate the development of novel approaches to AI ethics and responsible research and innovation methodologies more broadly, as it relies on an iterative design-implementation-validation cycle across increasing levels of autonomy, which, we hope, will enhance our ability to manage the ethical risks and societal impacts associated with future AI systems.

The Road Ahead

Our vision for incremental autonomy constitutes an attempt to develop a new foundation for engineering AI systems that can help crystallise recent advances and help overcoming the limitations and risks of present-day AI. It is underpinned by a focus on autonomy, which we view as essential to steer future research toward developing new AI capabilities that solve complex problems in real-world contexts, while, at the same time, favouring an incremental and integrative approach that will help build future complex capabilities on top of existing methods and solutions. Our iterative methodology, which emphasises the development of new methods to support developers in reusing, validating, and recombining components, is aimed at safeguarding human control and enabling a responsible approach to the development of future autonomous systems in the face of ever-increasing system complexity. We argue that this approach will also bring benefits in terms of improved sustainability, risk management, responsible innovation, as well as our ability to anticipate the social impact of AI.

Whether this vision will act as a positive force on AI communities and inspire new research initiatives and technological advances will of course depend on whether it can gain traction among these communities. While we hope that it will act as an interface between many different existing efforts and will contribute to important discussions about the future of AI research and its use in society, it will be necessary to engage in extensive discussions with researchers, practitioners, and users to gauge interest in pursuing this overall direction and, following this, a develop a more concrete research roadmap.