Genetic Programming and Evolvable Machines (2021) 22:137–139
https://doi.org/10.1007/s10710-020-09394-1

Virginia Dignum: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way

Springer Nature Switzerland AG, 2019. ISBN 978-3-030-30370-9 ISBN 978-3-030-30371-6

Nicolas E. Gold[1]

Received: 28 September 2020 / Accepted: 28 September 2020 / Published online: 10 October 2020

Artificial Intelligence (AI) is undoubtably a key technology of our contemporary age and one that affects all of us to a greater or lesser extent. Virginia Dignum's "Responsible Artificial Intelligence" addresses this topic in a relatively slim volume whose size belies the depth, breadth, quality, and clarity of the content. Central to the book's thesis is the responsibility on all those creating, designing, procuring, adopting, using, and subject to AI systems, to engage with each other in these activities. The treatment of these issues is simultaneously theoretical, philosophical, and practical, and leaves one in no doubt of the importance of treating AI technology and endeavour from a socio-cognitive-technical perspective.

The first chapter lays the ground for what follows, motivating the need for responsibility in AI and that this is different (although related) to considerations of ethics in AI. Specifically, Dignum highlights the action component of responsibility: that it is not sufficient simply to hold ethical views on AI, but one must act on them as well. This point is made clearly and effectively in the chapter, and the nature of responsibility (and indeed, the many different actors who hold that responsibility) is described at greater length. The need for new and ambitious governance is discussed, positioning AI systems very clearly as artefacts created to achieve goals in the service of humanity and thus subordinate to the contexts in which they exist. In positioning AI in this way, Dignum rightly identifies the transdisciplinary nature of AI and the need for contributions from social sciences, law, economics, cognitive sciences, and the humanities, in addition to engineering-centric technological advances. The overall thrust of the chapter is a nice balance between caution and optimism: caution about the potential risks of not considering the responsibilities involved in producing AI, and optimism for the potential benefits of responsibly-produced AI. Overall, the reader is well-prepared for the subsequent chapters through

✉ Nicolas E. Gold
n.gold@ucl.ac.uk

[1] University College London, London, UK

the establishment of a tripartite framework of application for ethics to AI: Ethics in, by, and for Design.

Chapters Two and Three examine the nature of AI to establish its roots and supporting disciplines, and then consider ethical decision-making. Engineering and philosophical views (among others) are covered. Properties of agency are discussed, followed by presentations of the nature of, and issues arising in, machine learning and interaction. Ethical decision-making is discussed with a brief introduction to key underlying theories of ethics and placement of these into practical AI-related contexts. The final part of the chapter presents approaches to the implementation of ethical reasoning. As in other chapters in the book, there are plentiful citations and recommendations of further sources for the interested reader.

Chapter Four presents a design methodology for the development of AI in accordance with principles of accountability, responsibility, and transparency. It contains helpful practical guidance but does not shy away from the complexity of the task. Clear definitions of the relevant concepts are offered and then applied. Of particular note is the emphasis on ensuring a broad values-driven approach that involves a wide range of stakeholders rather than a narrow focus solely on system performance.

The focus then shifts in Chapter Five to a consideration of whether, and how, an AI system could be developed that could reason ethically about its decisions and their consequences. Three broad approaches are considered and reflected upon, and common questions and issues identified (perhaps most importantly the question of whether such a system should be developed at all). There is a helpful discussion of the advantages and disadvantages of various approaches to value setting. The chapter concludes by examining ethical deliberation strategies, levels of behaviour, and the philosophical issues involved in considering whether AI systems have their own ethical status. It is here that Dignum makes a strong argument for the need to study autonomy from a distributed and socio-cognitive-technical perspective rather than considering AI systems as stand-alone entities.

Chapter Six deals with the governance of AI and how responsible AI can be put into practice. It clearly defines the differences between concepts such as regulation and certification, and relates responsible AI to codes of professional conduct (including aspects of training researchers and developers in their responsibilities to consider ethical aspects of their work). A large number of pointers to contemporary codes and discussions of AI ethics and responsibility are helpfully included.

Chapter Seven looks toward the future, considering the impacts of AI on jobs, education, and other societal areas, finally assessing the potential for super-intelligence.

The reader emerges from the book with an understanding of the importance of responsible AI development and use, a grasp of the breadth of issues involved, enough depth to know where to seek further information but not so much as to feel overwhelmed, and practical approaches to start adopting responsible AI perspectives in practice. The type of material in the book is well-balanced and even more so the argument. Dignum strikes a helpful tone of optimistic realism: the potential benefits of AI are strongly acknowledged but not hyped, the potential risks likewise clearly articulated, but the technology not written off. Overall one is left with the impression of a field in which technical advances are impressive but are not yet wholly matched by similar advances in understanding, regulation, and responsible development and use (but that this can be remedied).

The book claims to be suitable for undergraduate students, and interested and concerned researchers, practitioners, and citizens. It will serve these audiences well but would be of equal benefit and importance to policymakers and those making decisions about adopting AI technology who might otherwise be focused on a narrower functional perspective. Dignum takes a holistic and accessible approach to the material that makes it suitable for a wide range of audiences and that encourages more than one reading. Highly recommended for all.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.