111 / 2022, 4, Downloaded from https://online.library.wije.com/doi/10.1111/2ygo.12830 by Otto-Friedrich-Universität, Wiley Online Library on [03/02/2023]. See the Terms and Conditions (https://online.library.wije.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles as a governed by the applicable Creative Commons. Leieners and Conditions (https://online.library.wije.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles as a governed by the applicable Creative Commons. Leieners and Conditions (https://online.library.wije.com/terms-and-conditions) on Wiley Online Library.

# Artificial Intelligence and Religion: Recent Advances and Future Directions

with Andrea Vestrucci, "Introduction: Five Steps Towards a Religion—AI Dialogue"; Lluís Oviedo, "AI and Theology: Looking for a Positive—But Not Uncritical—Reception"; Christoph Benzmüller, "Symbolic AI and Gödel's Ontological Argument"; Sara Lumbreras, "Lessons from the Quest for Artificial Consciousness: The Emergence Criterion, Insight-Oriented AI, and Imago Dei"; Marius Dorobantu, "Artificial Intelligence as a Testing Ground for Key Theological Questions"; and Andrea Vestrucci, "Artificial Intelligence and God's Existence: Intersecting Theology and Computation."

## SYMBOLIC AI AND GÖDEL'S ONTOLOGICAL ARGUMENT

by Christoph Benzmüller

Abstract. Over the past decade, variants of Gödel's ontological arguments have been critically examined using modern symbolic AI technology. Computers have unearthed new insights about them and even contributed to the exploration of new, simplified variants of the argument, which now need to be further investigated by theologians and philosophers. In this article, I provide a brief, informal overview of these contributions and engage in a discussion of the possible future role of AI technology for the rigorous assessment of arguments in theology and philosophy.

Keywords: artificial intelligence; information; interdisciplinarity; metaphysics; ontology

### ARTIFICIAL INTELLIGENCE (AI): SYMBOLIC VERSUS SUBSYMBOLIC

The field of AI has already experienced various winter and summer periods in its comparatively young history. The current "heat wave" seems to be stronger and more sustained than previous summer periods, as significant progress has been made in the last two decades, especially in the area of *subsymbolic AI*. Data-driven machine learning methods, particularly *deep learning*, have been hyped in the media, and industry is desperately seeking and hiring experts, since AI is widely considered the steam engine of the twenty-first century. These success stories have even led to the term AI

Christoph Benzmüller is Full Professor for AI Systems Engineering at the Otto-Friedrich-University Bamberg, Bamberg, Germany; e-mail: christoph-benzmueller@uni-bamberg.de (or c.benzmueller@gmail.com).

[Zygon, vol. 57, no. 4 (December 2022)]

www.wileyonlinelibrary.com/journal/zygon

being largely reduced to subsymbolic AI in the public perception today. At the same time, *symbolic AI* is now often referred to as *good old fashioned AI* (GOFAI), which is clearly a misleading term, as both symbolic and subsymbolic AI techniques have been explored as relevant from the beginning, particularly with respect to visions of strong AI. And as will be discussed below, there are success stories to report in the area of symbolic AI as well, although not at the level of subsymbolic AI, which currently promises to be a very robust and viable choice for many low-hanging fruit applications in industry.

The vision of *strong AI*, that is, AI that surpasses human capabilities in all or almost all domains, requires, in my view, the hybridization of techniques from both sides (or a convincing explanation of why symbolic reasoning techniques should suddenly and readily evolve from data-driven subsymbolic techniques). It is thus in fact the area of hybrid AI where the next "big (intellectual) thing" is to expected. To better understand my viewpoint and to see why I continue to insist on the relevance of symbolic AI, it appears useful that I briefly present my personal working definition of the term AI:

AI is a science of computer technologies developed to achieve and explain *intelligent* behavior in machines. By *intelligent* I refer here to a collection of abilities that enable an entity to (i) solve or learn how to solve certain (difficult) domain specific problems, (ii) master the known and the unknown, i.e., to act successfully in known, unknown, and dynamic environments (which requires perception, planning, agency, etc.), (iii) think rationally, avoid contradictions, and explore abstract theories, (iv) self-reflect, recognize self-contradictions, and reconcile one's reasoning with higher-level goals and norms, and (v) interact socially with other entities and reconcile one's goals and norms with those of a community (for a greater good).

Excellent progress in the field of AI has been made primarily at level (i) and to some extent at level (ii); this progress at both levels has been enabled by both subsymbolic and symbolic techniques, although the emphasis at present, especially at level (i), seems to be more on the former. Progress at level (iii), in my view, clearly requires symbolic modeling and reasoning skills in an intelligent entity; in particular, the exploration and subsequent evaluation of an abstract theory, say in mathematics, traditional natural sciences, or metaphysics, inevitably involves mastering a *symbolic representation language*. Bibel (2022) has therefore introduced the term "representing objects" to refer to the elements of this language, and he argues that with the invention of the computer a "new natural science" has emerged, capable of theorizing about representing objects that are enabled by AI as scientific discipline (including, in particular, symbolic AI) and which is way more than just a set a of, for example, machine learning tools.

Moreover, in mathematics and metaphysics in particular, abstract theories are not seldom explored using *very little or no experimental data*. An

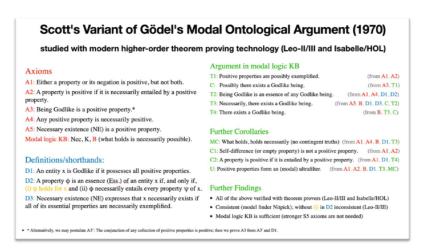
example is the ontological argument for the existence of God, since it is based entirely on concepts and definitions explored by abstract thought. Among the several kinds of ontological arguments, the one formulated in modern modal logic by Kurt Gödel and subsequently modified by Dana Scott and others has been the focus of some of my own work with colleagues over the past decade. This research has been conducted with fruitful humanmachine interaction by utilizing symbolic representation and communication techniques, and it has now led to the exploration of a highly simplified theory, which can, in turn, further inform human insight into the ontological argument. While so far only parts of the exploration process have been fully automated by symbolic AI reasoners (automated theorem provers and model finders), there are indications that the pendulum here may swing more strongly toward such symbolic AI reasoners and increased automation in the future (especially when future hybrid AI systems will be used). In purely subsymbolic AI systems, lacking any symbolic representation means, neither abstract theory exploration nor subsequent theory evaluation or adequate communication with humans seems sufficiently well supported to enable fruitful theory exploration. In other words, subsymbolic AI skills will hardly be sufficient to automate abstract theory exploration processes in intelligent agents, since such processes seem inherently dependent on formal (symbolic) language skills.

The computer-supported experiment on the ontological argument reported in the next section has thus a twofold function in this article. First, it serves as one, small piece of evidence for recent advances in the field of knowledge representation and reasoning (KR&R), that is, symbolic AI, which has led to new insights that are relevant for philosophy and rational theology. Second, it provides evidence for the claim that the creation, use, and manipulation of explicit symbolic representations is central to certain scientific discovery processes. In other words, a strong AI lacking explicit KR&R skills appears essentially inconceivable to me.

### A Computationally Explored Simplified Variant of the Ontological Argument

What is the simplified ontological argument, resp. simplified ontological theory, the computer helped to explore? I switch from "argument" to "theory" here, since not only the pure argument is of interest for our investigation, but also some further implications arising from the axioms and definitions of the argument (cf. the modal collapse [MC] below).

This new theory resulted from a series of experiments in which Scott's (1972) variant of Gödel's (1970) theory has been taken as the starting point of investigation. In Figure 1, I briefly recap Scott's theory before I present the new simplified theory.



**Figure 1.** Experiments on Scott's Variant of Gödel's Modal Ontological Argument. [Color figure can be viewed at wileyonlinelibrary.com]

The notion of positive properties is taken by both Gödel and Scott as the single primitive notion, based on which the concept of a Godlike entity is then defined by (definition D1) "A Godlike entity possesses all positive properties." Axiom A1 states that "Either a property or its negation is positive, but not both," and Axiom A2 expresses "A property is positive if it is necessarily entailed by a positive property"; property entailment is defined here using material implication. Axiom A3, in Scott's variant,<sup>2</sup> simply postulates that "Being Godlike is a positive property," and Axiom A4 that "Any positive property is necessarily positive." Axiom A5 finally states that "Necessary existence is a positive property," where "x necessarily exists if all of its essential properties are necessarily exemplified" (definition D3); definition D3 in turn uses the term "essential properties," resp. "essence," which is defined by "A property  $\varphi$  is an essence of an entity x if, and only if, (i)  $\varphi$  holds for x and (ii)  $\varphi$  necessarily entails every property  $\psi$  of x" (definition D2). The clause (i) in D2 is omitted in Gödel's (1970) variant and was added by Scott (because he found it an intuitive and natural to do so).

The above axioms and definitions of Scott's theory variant were in a first round of experiments formalized in a higher order modal logic S5,<sup>3</sup> whereas the logic S5 was itself introduced, utilizing the logic-pluralistic KR&R methodology termed LogiKEy (Benzmüller et al. 2020), as a theory of classical higher-order logic (HOL). Due to this nested theory construction, off-the-shelf interactive and automated theorem provers for HOL became readily applicable to support an in-depth formal assessment of Scott's ontological theory with computers.

The conducted experiments quickly confirmed the validity of each of the following argument steps; thereby the theorem provers automatically identified the minimal dependencies as shown in brackets; from these assumptions the statements follow:

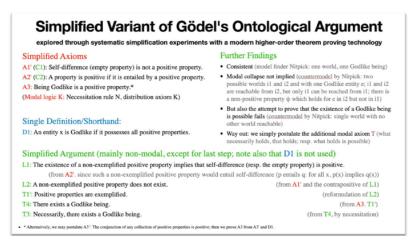
- T1: "Positive properties are possibly exemplified" (from A1, A2).
- C: "Possibly there exists a Godlike being" (from A3, T1).
- T2: "Being Godlike is an essence of any Godlike being" (from A1, A4, D1, D2).
- T3: "Necessarily, there exists a Godlike being" (from A5, B, D1, D3, C, T2).
- T4: "There exists a Godlike being" (from B, T3, C).

Subsequently, further implied theorems were explored, including:

- MC: "What holds, holds necessarily (no contingent truths)" (from A1, A4, B, D1, T3).
- C1: "Self-difference (or empty property) is not a positive property" (from A1, A2)
- C2: "A property is positive if it is entailed by a positive property" (from A1, D1, T4)
- U: "Positive properties form a (modal) ultrafilter" (from A1, A2, B, D1, T3, MC)

C1 and C2 are two of the filter/ultrafilter properties that were tested in step U of the experiments; the reason they are additionally listed here, resp. singled out, will become clear below.<sup>4</sup>

All the above argument steps were verified by Benzmüller and Woltzenlogel Paleo (2014, 2015, 2016a) with the Leo automated theorem provers (Benzmüller et al. 2015; Steen and Benzmüller 2021) for HOL and later also with the interactive proof assistants Isabelle/HOL (Nipkow et al. 2002) and Coq (Bertot and Casteran 2004). In addition, the Nitpick model finder (Blanchette and Nipkow 2010) for HOL was used to automatically confirm the consistency of the theory; it provided a simple model. However, when clause (i) in definition D2 was omitted in the experiments, no such model was found anymore, suggesting an attempt to infer falsity from the axioms of the theory. The theorem prover Leo-II actually succeeded in this and computed a resolution-style proof, which, with some delay, was analyzed and converted by Benzmüller and Woltzenlogel Paleo (2016a, 2016b) into a human-readable intuitive proof argument. This first round of experiments with the modal variants of Gödel and Scott's ontological argument was thus very successful. The first variant was shown to be inconsistent and the second variant valid, and for this the



**Figure 2.** Simplified Variant of Gödel's Ontological Argument; Explored in a Human-Machine Interaction. [Color figure can be viewed at wileyonlinelibrary.com]

theorem provers needed only the modal logic KB and not the modal logic S5, as it is had been controversially discussed in this context.

These successful experiments were then later taken as a starting point by Fuenmayor and Benzmüller (2017) and Benzmüller and Fuenmayor (1990) not only for the verification of Fitting's (2002) modal variant but also for an extended, comparative study of Fitting's with the modal variants of Anderson (1990), resp. Anderson and Gettings (1996). In these experiments, it was shown that both Fitting's and Anderson's variants were also both valid and that they indeed, as intended by their authors, avoid the controversial MC. While both variants are seemingly different, they are nevertheless related when comparing these variants through "ultrafilter lenses"; they apply related relaxations regarding the concrete ultrafilter properties that must be satisfied by the set of positive properties (more precisely, the set of modally rigidified positive properties).

The ultrafilter perspective was then also the trigger for the investigation of the strongly simplified variant presented in Figure 2. Thereby, the two special ultrafilter properties C1 and C2 from above proved to be especially interesting, because, as the computer automatically confirmed, the existence (resp. necessary existence) of a God-like entity can already be derived very easily in the basic modal logic K (resp. in the modal logic KT) from these two corollaries of Scott's theory in combination with the Axiom A3.

The new simplified theory starts out with unmodified definition of a Godlike entity as an entity that possess all positive properties, in which the notion of "positive property" is the only primitive concept that is further constrained by the following three axioms:

- C1: "Self-difference, or, alternatively, the empty property, is not a positive property."
- C2: "A property is positive if it is entailed by a positive property."
- A3: "Being Godlike is a positive property."

The assumed logic for the moment is a modal logic K. It follows:

- L1: "The existence of a non-exemplified positive property implies that self-difference (resp. the empty property) is positive" (from A2', since such a nonexemplified positive property would entail self-difference, where p entails q is defined as: for all x, p(x) implies q(x)).
- L2: "A non-exemplified positive property does not exist" (from A1' and the contrapositive of L1).
- T1': "Positive properties are exemplified" (a simple reformulation of L2).
- T4: "There exists a Godlike being" (follows from A3 and T1').
- T3: "Necessarily, there exists a Godlike being" (from T4, by necessitation).

Interestingly, the possibility of the existence of a God-like being cannot be easily proved in the basic modal logic K, and Nitpick reports a countermodel. However, this circumstance can be countered simply by switching to modal logic KT, that is, by adding the modal axiom T (the T axiom states, "What necessarily holds, holds" or, alternatively, "What holds, is possible"). Moreover, in modal logic KT also the following alternative proof variant was automatically confirmed by the theorem provers integrated with the Isabelle/HOL proof assistant; this variant is close to the variant by Scott as shown at the beginning:

- T1: "Positive properties are possibly exemplified" (from A1', A2', T).
- C: "Possibly there exists a Godlike being" (from A3, T1).
- T2': "The possible existence of a Godlike being implies its necessary existence" (from A3, A1', A2').
- T3: "Necessarily, there exists a Godlike being" (from T2', C).
- T4: "There exists a Godlike being" (from T3, T).

Future work includes the computational study of the relationship of the above human-explored simplified variants of the modal ontological argument to other variants of the argument that are still less known in the literature, such as Gustaffson (2020), Gödel's variant #2 discussed by Kanckos and Lethen (2021), and the proposals by Świetorzecka and Łyczak (2020) and Christian (1989).

960 Zygon

### Metaphysics and Rational Theology as an Experimental Science?

As the above experiments show, modern symbolic AI technology can thus make a fruitful contribution to the rigorous evaluation of existing arguments and theories in metaphysics and rational theology, and even support the exploration of new, simplified theories. Thanks to the LogiKEy methodology, even the object logics in which the domain theories of interest are modeled and encoded become negotiable (i.e., subject to modification, extension, or replacement) and thus objects of study in the computer. Currently, however, a collaboration triangle is still required, involving a domain expert (e.g., philosopher or theologian), an AI expert/logician (who is proficient in the automated and interactive reasoners and the LogiKEy methodology), and the AI reasoners. In the future, the AI expert/logician should ideally become less and less relevant as AI reasoners continue to improve both proof automation and intuitive user interaction. Assuming a strong AI vision, another interesting question arises. Namely, whether an AI reasoner initialized with an appropriately represented object logic, a domain theory, and a proof argument as a starting point can eventually perform a creative process as the one described in the section "A Computationally Explored Simplified Variant of the Ontological Argument," but completely autonomously.

After hours of theorizing, in which the AI continuously manipulates and (re)assesses these representing objects, it would then, for example, inform humans of the most interesting simplified or new alternative theories (and their assumed object logics) that it has explored along the way, or it would return interesting countermodels that it has found. In the spirit of our initial discussion, such a vision could then even be seen as a shift from metaphysics and rational theology toward an experimental, natural science in which representing objects of these domains are physically manipulated and assessed in the computer. This is, of course, a controversial vision that in itself deserves further analysis.

In summary, symbolic AI can do much for metaphysics and rational theology. Conversely, metaphysics and rational theology are of particular interest to AI research, especially because of the prominent role that thought experiments and abstract arguments play in these disciplines in combination with the typical absence of "real world" data. These disciplines therefore particularly emphasize the need for explicit symbolic representations and theorizing with little or no data.

#### Notes

1. Since its very beginnings, the field of AI has differentiated between the *connection-ist/subsymbolic* and the *symbolic* paradigms for modelling and explaining intelligent behavior. *Subsymbolic* approaches generally model intellectual abilities using artificial neural networks,

that is, networks of computing units devoid of semantic meaning. On the other hand, the *symbolic* approach assumes that intelligence results from the manipulation of abstract compositional and meaningful representations. Techniques used in this field include rule-based systems and formal logic. Both paradigms have well-known strengths and weaknesses, and the debate about whether human-level intelligence can be plausibly modeled and explained by one or the other has a long tradition.

- 2. In Gödel's scriptum we find the more complex (third-order) axiom: "The conjunction of any collection of positive properties is positive"; let's call this A3\*. Using D1, it can be shown that A3 is a consequence of A3\* (in modal logic K) and, moreover, A3 suffices for establishing the intended results. Scott thus found it reasonable to replace A3\* by the much simpler axiom A3
- 3. And additionally a logic S5U was used in the experiments by Benzmüller and Woltzenlogel Paleo (2016a, b).
- 4. An ultrafilter U (on sets) is a filter satisfying an additional maximality criterion (for each property  $\varphi$  either  $\varphi$  or its complement is an element of U). And a filter U is a set of properties, which is large (i.e., the universal property is an element of U), which excludes the empty property (cf. C1) and which is closed under supersets (cf. C2) and intersections. Principle ultrafilters have a smallest element, which satisfies all properties of the ultrafilter; such a smallest element in our context is associated with the concept/property of a God-like being. It is relevant to remark that the notion of filter/ultrafilter needed to be adapted for the modal context of the ontological argument; for more details, see the discussion in Benzmüller and Fuenmayor (2020); in the latter work is has been shown how seemingly rather different variants of the modal ontological argument are nevertheless closely related from the perspective of the ultrafilter structures they axiomatize.

#### References

- Anderson, Anthony C. 1990. "Some Emendations of Gödel's Ontological Proof." Faith and Philosophy 7 (3): 291–303. https://doi.org/10.5840/faithphil19907325
- Anderson, Anthony C., and Michael Gettings. 1996. "Gödel's Ontological Proof Revisited." "Gödel'96: Logical Foundations of Mathematics." *Computer Science, and Physics: Lecture Notes in Logic* 6:167–72. Springer. https://doi.org/10.1017/9781316716939.011
- Benzmüller, Christoph, and David Fuenmayor. 2020. "Computer-Supported Analysis of Positive Properties, Ultrafilters and Modal Collapse in Variants of Gödel's Ontological Argument." Bulletin of the Section of Logic 49 (2): 127–48. https://doi.org/10.18778/0138-0680.2020.08
- Benzmüller, Christoph, Xavier Parent, and Leon van der Torre. 2020. "Designing Normative Theories for Ethical and Legal Reasoning: LogiKEy Framework, Methodology, and Tool Support." *Artificial Intelligence* 287:103348. https://doi.org/10.1016/j.artint.2020. 103348
- Benzmüller, Christoph, Nik Sultana, Lawrence C. Paulson, Frank Theiss, et al. 2015. "The higher-order prover LEO-II." *Journal of Automated Reasoning* 55 (4): 389–404. https://doi.org/10.1007/s10817-015-9348-y
- Benzmüller, Christoph, and Bruno Woltzenlogel Paleo. 2014. "Automating Gödel's Ontological Proof of God's Existence with Higher-Order Automated Theorem Provers." *ECAI 2014, Frontiers in Artificial Intelligence and Applications* 263:93–98. IOS Press. https://doi.org/10.3233/978-1-61499-419-0-93
- 2015. "Interacting with Modal Logics in the Coq Proof Assistant." In Proceedings of Computer Science—Theory and Applications—10th International Computer Science Symposium in Russia, CSR 2015, edited by L. D. Beklemishev and D. V. Musatov, 398–411. Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-3-319-20297-625
- 2016a. "The Inconsistency in Gödel's Ontological Argument: A Success Story for AI in Metaphysics." In *IJCAI 2016*, edited by S. Kambhampati, 936–42. Menlo Park, CA: AAAI Press/The MIT Press. https://www.ijcai.org/Proceedings/16/Papers/137.pdf
- 2016b. "An Object-Logic Explanation for the Inconsistency in Gödel's Ontological Theory (Extended Abstract, Sister Conferences)." In KI 2016: Advances in Artificial

- Intelligence, Proceedings, edited by M. Helmert and F. Wotawa, 244–50. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-46073-4
- Bertot, Yves, and Pierre Casteran. 2004. "Interactive Theorem Proving and Program Development—Coq'Art: The Calculus of Inductive Constructions." *Texts in Theoretical Computer Science. An EATCS Series*. Berlin: Springer. https://doi.org/10.1007/978-3-662-07964-5.
- Bibel, Wolfgang. 2022. Komputer kreiert Wissenschaft. Informatik Spektrum. https://doi.org/10. 1007/s00287-022-01456-1
- Blanchette, Jasmin C., and Tobias Nipkow. 2010. "Nitpick: A Counterexample Generator for Higher-Order Logic Based on a Relational Model Finder." In *Interactive Theo*rem Proving—ITP 2010, edited by M. Kaufmann and Lawrence C. Paulson, 131– 46. Springer, Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-14052-5 11
- Christian, Curt. 1989. "Gödel Version des Ontologischen Gottesbeweises." Sitzungsberichte der Österreichischen Akademie der Wissenschaften, Abt. II 198:1–26.
- Fitting, Melvin. 2002. *Types, Tableaus, and Gödel's God.* Springer Dordrecht: Kluwer. https://doi.org/10.1007/978-94-010-0411-4
- Fuenmayor, David, and Christoph Benzmüller 2017. "Automating Emendations of the Ontological Argument in Intensional Higher-Order Modal Logic." KI 2017: Advances in Artificial Intelligence, Proceedings, edited by G. Kern-Isberner, J. Fürnkranz and M. Thimm, 114–27. Springer, Cham. https://doi.org/10.1007/978-3-319-67190-1\_9
- Gödel, Kurt. 1970. Appendix A. Notes in Kurt Gödel's Hand. In Sobel, J. 2004. Logic and Theism: Arguments for and Against Beliefs in God. Cambridge University Press. 144–5. https://doi.org/10.1017/CBO9780511497988.
- Gustaffson, Johan E. 2020. "A Patch to the Possibility Part of Gödel's Ontological Proof." *Analysis* 80 (2): 229–40. https://doi.org/10.1093/analys/anz024
- Kanckos, Annika, and T. Lethen 2021. "The Development of Gödel's Ontological Proof." The Review of Symbolic Logic 14 (4): 1011–29. https://doi.org/10.1017/ S1755020319000479
- Nipkow, Tobias, Lawrence C. Paulson, and Makarius Wenzel 2002. Isabelle/HOL—A Proof Assistant for Higher-Order Logic. Springer Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-45949-9
- Scott, Dana S. 1972. Appendix B: Notes in Dana Scott's Hand. In Sobel, J. 2004. Logic and Theism: Arguments for and Against Beliefs in God. 145–6. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511497988
- Steen, Alexander, and Christoph Benzmüller 2021. "Extensional Higher-Order Paramodulation in Leo-III." *Journal of Automated Reasoning* 65:775–807. https://doi.org/10.1007/s10817-021-09588-x
- Świetorzecka, Kordula, and Marcin Łyczak. 2020. "An Even More Leibnizian Version of Gödel's Ontological Argument." In *Beyond Faith and Rationality. Sophia Studies in Cross-Cultural Philosophy of Traditions and Cultures*, edited by R. S. Silvestre, B. P. Göcke, J. Y. Béziau, and P. Bilimoria, 34. Cham, Switzerland: Springer. https://doi-org.libproxy.viko.lt/10. 1007/978-3-030-43535-6\_6