How the curse of intractability can be cognitive science's blessing

Iris van Rooij (i.vanrooij@donders.ru.nl)
Radboud University Nijmegen, Donders Institute for Brain, Cognition, and Behaviour
Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

Keywords: philosophy of cognitive science; computational explanation; intractability; computational complexity; inter-theory reduction

Intractability has been a thorny issue in cognitive science. Informally, intractability refers to the problem that computations that work for small toy domains cannot scale to domains of real-world complexity due to prohibitive resource consumption. It is not uncommon for cognitive scientists to try to discredit theories in competing frameworks by pointing out that those frameworks run into intractability issues. For instance, both connectionists and Bayesians have argued against symbolic and logic approaches, respectively, because the latter two would yield intractable theories of cognition (Haselager, 1997; Oaksford & Chater, 1998). Yet, it is now well known that also connectionist and Bayesian theories can be intractable (Frank, Haselager, & van Rooij, 2009; Kwisthout, Wareham, & van Rooij, 2011). Moreover, even heuristic and dynamical systems theories, both often lauded for being tractable, seem unable to live up to that image when forced to scale beyond toy domains (van Rooij, Wright, & Wareham, 2012; van Rooij, 2012). Evidently, intractability is not a problem for specific theories, or even for specific theoretical frameworks. Instead, it seems a ubiquitous feature of theoretical frameworks with high degrees of generality.

In this talk, I put forth the argument that cognitive science may have been too quick in seeing intractability as "just bad news" and that the field has been losing out on the opportunity to turn what seems to be a curse into a blessing. The upshot of my argument will be that the ubiquity of intractability can better be seen as a useful theoretical guide to the boundaries of cognition's domain-generality. Using formal notions from computational complexity theory, I will explain how intractability helps demarcate those boundaries.

Computational-level theories

Theories of cognition can be formulated at different levels of explanation. For instance, following Marr's (1982) widely used tripartite distinction, a theory can be formulated at the computational level ('what is the computational problem (or function) being computed?'), the algorithmic level ('what algorithm is used for the computation'), or the implementational level ('how is the algorithm physically realised?'). Here, I will focus on computational-level theories.

Formally, a computational level theory can be conceived of as a mathematical function, $T : I_T \rightarrow O_T$, mapping inputs $i \in I_T$ to outputs $T(i) \in O_T$. Defining a computational-level theory T of some cognitive capacity ϕ involves defining both the input domain I_T and output domain O_T and the nature of the mapping, $T : I_T \rightarrow O_T$, between them. We will say that T is an *accurate* characterisation of a cognitive capacity $\phi : I_\phi \rightarrow O_\phi$ if and only the following three conditions are met:

1. $T(i) = \phi(i)$, for every $i \in I_\phi$
2. $T(i) = \phi(i)$, for every $i \in I_T$
3. $I_T = I_\phi$

It may happen that condition (1) is met, without conditions (2) and (3) being met. This can happen, for instance, when $I_T \supset I_\phi$. In that case, we say that T is an *overgeneralisation* of ϕ. Note that an overgeneralization describes ϕ accurately for inputs confined to $I_\phi \subset I_T$, but that I_T includes inputs in its domain that are outside the scope of the capacity ϕ. Conversely, it can also happen that condition (2) is met, without conditions (1) and (3) being met. This can happen, for instance, when $I_T \subset I_\phi$. In that case, we say that T is an *undergeneralisation* of ϕ. Note that an undergeneralization describes ϕ accurately for inputs confined to $I_T \subset I_\phi$, but that I_T fails to include inputs in its domain that are within the scope of the capacity ϕ. In cases where neither condition (1) or (2) is met, we consider T to be a mischaracterisation of ϕ.[1]

Coming up with accurate characterisations of cognitive capacities is no easy task, given the sheer size of the space of possible distractors: For any given capacity ϕ there exist in principle infinitely many possible mischaracterisations, many of which may even seem plausible in light of existing empirical observations. This underdetermination problem has long been known and has motivated theorists to propose theoretical constraints on candidates for T, such as rationality (Anderson, 1990), or tractability (van Rooij, 2008), or a combination of the two (van Rooij, Wright, Kwisthout, & Wareham, 2014).

Intractability and NP-hardness

Although there are many notions of intractability of relevance for cognitive science, here we consider specifically the notion of NP-hardness (Arora & Barak, 2009). If a computational-level function $T : I_T \rightarrow O_T$ is NP-hard, then all algorithms computing T require super-polynomial time for some (infinitely many) inputs in the domain I_T.[2] To illustrate why super-polynomial time algorithms are considered intractable, let us take an exponential-time algorithm (taking on the order of 2^n time) as an example. Even if we were to assume that the algorithm could perform 1000 computation steps per

[1] Arguably, mischaracterisations come in different kinds and degrees, but for purposes of the argument I put forth here we need not concern ourselves with such details, as nothing in the argument hinges on them.

[2] This is assuming P $\neq$ NP, a conjecture generally believed by computer scientists and uncontested by cognitive scientists.

second—say, in parallel—the algorithm would run for 12 days on an input size of $n = 30$, and 35 millennia for an input size of $n = 50$. It is because of such prohibitive resource demands that super-polynomial time algorithms are generally considered unfeasible for anything but small inputs. Given that many cognitive capacities operate on inputs of intermediate to large sizes (e.g., vision, categorisation, language learning, belief fixation), super-polynomial time algorithms are generally computationally implausible. Hence, the same holds for NP-hard computational-level theories.

Inter-theory reductions and domain generality

As alluded to above, many computational-level theories in cognitive science are NP-hard. This intractability is not specific to any particular type of model, nor is it specific to a particular cognitive capacity, as intractability is observed for models of perception, language, reasoning, categorisation, decision making, and motor planning (Kwisthout, et al., 2011; van Rooij & Wareham, 2012). This ubiquity of NP-hardness can be understood as a natural consequence of attempting to scale one's computational-level models to general domains, without regards for tractability. To explain why this is so, we will consider computational-level theories in terms of their inter-theory subsumption and reduction relations (see Figure 1).

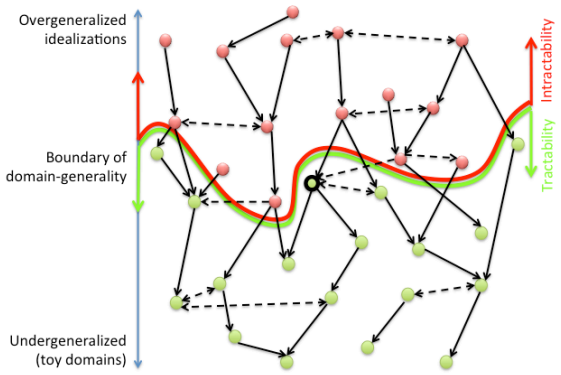

Figure 1: Illustration of hypothetical space of possible theories (circles) for a given capacity ϕ. See text for details.

Let us say that a computational-level theory $T : I_T \rightarrow O_T$ *subsumes* another $T' : I'_T \rightarrow O'_T$ whenever $T(i) = T'(i)$ for all $i \in I_T$ and $I_{T'} \subset I_T$. In such case, we also say T' *reduces* to T.[3] It is known that tractability is inherited along the direction of the subsumption relation, and intractability is inherited along the direction of the reduction relation. Put differently, if a theory T is tractable, then so are all theories that it subsumes; and if a theory T is intractable then so are all theories that it reduces to.

In Figure 1 subsumption relations between theories (circles) are denoted by solid arrows, with the reduction relation running in the opposite direction. Other (less direct) forms of inter-theory reductions are depicted by dotted arrows (see footnote 3). For the sake of argument, let the green thick circle denote an accurate computational-level characterisation T of ϕ. Note that T subsumes undergeneralisations of ϕ that belong to toy domains as well as reduces to overgeneralisations of ϕ that are intractable. Many accurate computational-level theories of relevance for cognitive science may similarly lie somewhere on a path crossing the boundary between tractable and intractable domains. This is to be expected, given that cognitive capacities require quite expressive formalisms for their accurate characterisation and expressive functions are typically intractable for unrestricted domains.

In this perspective, hitting upon an intractable characterisation T' does not mean that one has mischaracterised the cognitive capacity ϕ. It could simply mean that T' is an overgeneralisation. By exploring computational-level theories that are subsumed by T' one may identify several tractable, but still quite domain-general, candidates for an accurate computational-level characterisation. Such a strategy would also help map out the border between tractable and intractable computational-level theories for cognitive capacities, which can provide cognitive science with a useful view on what are the scope and limits of domain generality for resource-bounded cognition, be it human or artificial.

[3] The inverse of subsumption is a special case of the more general notion of polynomial-time reduction (Arora & Barak, 2009). With the latter, it can also be shown that theories in different frameworks are of comparable (in)tractability (dotted lines, Fig. 1).

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Arora, S., & Barak, B. (2009). *Computational complexity: a modern approach*. Cambridge University Press.

Frank, S. L., Haselager, W. F., & van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition*, *110*(3), 358–379.

Haselager, W. (1997). *Cognitive science and folk psychology: The right frame of mind*. Sage.

Kwisthout, J., Wareham, T., & van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, *35*(5), 779–784.

Marr, D. (1982). *Vision.* Freeman New York.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world.* PTaylor & Francis.

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive science*, *32*(6), 939–984.

van Rooij, I. (2012). Self-organization takes time too. *Topics in cognitive science*, *4*(1), 63–71.

van Rooij, I., & Wareham, T. (2012). Intractability and approximation of optimization theories of cognition. *Journal of Mathematical Psychology*, *56*(4), 232–247.

van Rooij, I., Wright, C. D., Kwisthout, J., & Wareham, T. (2014). Rational analysis, intractability, and the prospects of as if-explanations. *Synthese*, 1–20.

van Rooij, I., Wright, C. D., & Wareham, T. (2012). Intractability and the use of heuristics in psychological explanations. *Synthese*, *187*(2), 471–487.