The legally mandated approximate language about AI

Sjur Dyrkolbotn*[1] and Truls Pedersen†[2]

[1]Western Norway University of Applied Sciences
[2]University of Bergen

Abstract

In light of the current explosion of application of machine learning in data analysis and inference, we examine a particular challenge raised by the new EU General Data Protection Regulation (GDPR). The challenge we address pertains particularly to the demand that analyses of a person's data must be comprehensible to that person.

While there is a long tradition in viewing the world in terms of objects and properties in intuitive ways, recent decades have entertained a tension between more rule-based theories of mind (e.g., the representational theory of mind) and more holistic approaches (e.g., connectionism). While both approaches have merit, one seems to depart too much from a classical understanding of "knowing" to adequately satisfy the imminent legal realisty, and the other seems to be incapable of adequately capturing modern data analysis (as of yet).

As a solution to this predicament we propose a pragmatic compromise based on argumentation theory which seems to be able to provide a solid foundation in classical concepts, while at the same time permitting enthymematic presuppositions. We argue that developing a framework for explaining machine behavior in terms of abstract argumentation theory can address this dilemma.

1 Introduction

In light of the current explosion of application of machine learning in data analysis and inference, we examine a particular challenge raised by the new EU General Data Protection Regulartion (GDPR). Specifically, the GDPR (Sections 13-15) gives users targeted by automated decision-making a right to obtain "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject". What is the meaning and extent of this principle? This question has raised a lot of debate, with diverging views emerging as to how the new provisions should be interpreted.

*Sjur.Kristoffer.Dyrkolbotn@hvl.no
†Truls.Pedersen@uib.no

This paper was presented at the NIK-2018 conference; see http://www.nik.no/.

Some argue that the right to "meaningful information", taken in the context of the GDPR as a whole (especially in view of Section 22), establishes a *right to an explanation* of individual automated decisions [3, 10]. So, for instance, if John Doe is denied credit by a computer system that processes loan applications, John Doe will be entitled to an explanation of why the system rejected his application. By contrast, others have argued that the GDPR only gives data subjects a right to obtain a generic system description, providing some information about the "logic" generally used to reach decisions, without offering a concrete explanation for a given decision [11].

For rule-based symbolic systems, the distincton between an *ex ante* right to a system description and an *ex post* right to an explanation is important in practice, but not fundamental at the theoretical level. Rule-based systems admit *ex ante* descriptions (e.g., decision trees) that will – *in theory* – enable the data subject to predict what decisions will be made about him on the basis of the available data (which the data subject has an independent right to access). This, in turn, means that individual decisions can be "explained" *ex post*, by simply examining the data and the system description. For systems making use of machine learning techniques, this is different. For such systems, the link between *ex ante* descriptions of the system and *ex post* explanations of its decisions is inherently opaque. This is in part due to the fact that neither descriptions nor explanations are readily available for such systems, at least not when they are supposed to be provided in terms of "logic", as required by the GDPR. Plainly, we have no clear definition of what counts as an adequate system description for machine learning algorithms, much less an adequate explanation of individual decisions.

In the following, we discuss some challenges that arises from this, from the point of view of interpreting the legal language of the GDPR and giving meaningful theoretical content to it from a computer science perspective. We argue that focusing on the perceived distinction between *ex ante* and *ex post* forms of information is a misguided approach to the problem. As we have seen, neither persepective actually provides a well-defined approach to providing information about such systems, so the debate about which mode of explanation the GDRP requires seems rather beside the point. On the legal side, furthermore, the language of the GDPR is simply too vague to support any definite conclusion as to the extent of the data subjects' rights. This, we argue, is a *feature* of the text as a legal document, not a bug. It facilitates legal dynamism, whereby technology providers and courts must hear and consider diverging arguments about the scope of data subjects' rights in relation to automated decision-making. This will facilitate the development of an evolving legal standard, deepening our understanding of the legal implications of machine learning while helping to maintain congruence between law and technology.

The situation is similar from the perspective of computer science; the question of what counts as "meaningful information" about machine learning algorithms is a fundamental question, raising unresolved questions addressed in philosophy of mind and knowledge. Hence, it is unrealistic to expect any conclusive definition of what counts as meaningful information about machine learning, as much as it is unrealistic to expect computer scientists to settle deep questions of epistomology. We argue that the best way forward, in light of this, is to encourage argumentation and debate about automated decision-making, setting up a climate of critical inquiry that can produce "meaningful information" through a dialectical process that resembles the manner in which people inquire into the motives and reasons for human behaviours. The role of computer science, in this regard, is to facilitate argumentation that is sound, exploring the unknown without ignoring those basic

rules of argument and partial facts about machine learning that computer science can in fact provide and (possibly) enforce. However, while we must always aim for proper and correct explanations, we also have to ensure that the explanations are not so complex that *users* are unable to understand them.

2 Legal background

The basis of the right to "meaningful information about the logic involved" in automated decision-making is found in Articles 13-15 of the GDPR, which all provide for this right under different triggering conditions. While Articles 13 and 14 are triggered when data collection takes place, Article 15 provides a more general right of access to information for any subject whose personal information is "processed" by the system (including the right to know whether or not personal information is actually being processed).

Consequently, the right to "meaningful information" kicks in not only when personal data is collected, but also when it is used to inform an automated decision. This means that a right to explanation for individual decisions *ex post* can be plausibly suggested as one possible interpretation of the right to meaningful information. However, as argued in [11], the legislative process leading to the final version of the GDPR suggests that an explicit right to an explanation *ex post* was *intentionally left out* of the final version. This can suggest a more restrictive interpretation, but it is hardly a conclusive argument. It is perfectly possible, for instance, to regard an explicit right to an explanation as redundant and potentially misleading alongside a more general right to "meaningful information about the logic involved". It is worth noting, for instance, that a "right to explanation" of a decision is arguably far *less* demanding and *more* open to vacuous interpretations than a requirement that the subject must be supplied with "meaningful information about the logic".

To exemplify, if John Doe is denied credit, an explanation amounting to the fact that the system judged him to be a "high risk" applicant, on the basis of an "industry standard" credit rating algorithm, might be taken to satisfy a lax interpretation of an *ex post* explanation requirement. Indeed, reason-giving drawing on authority, reputation and experience (as in the "industry standard" reference) cannot be rejected when the requirement is simply to provide an explanation. However, when the requirement is to provide meaningful information about *the logic involved*, we are forced to reject explanations of this nature. In relation to the logic, appealing to authority and experience becomes a *fallacy of argumentation*. Ruling out this kind of argumentation is a feature of the language of the GDPR, which might have been diluted if the right to access the logic of the system was qualified by a more specific – but potentially less demanding – "right to an explanation" for individual decisions.

The broader point is that while the GDPR is open to interpretation and diverging arguments about the scope of data subjects' rights, it anchors those interpretations and arguments in a rather demanding criteria focusing on the logic involed in decision-making. For systems relying on machine learning, this anchor is *far more signficant* than the ambiguity of whether explanations must be provided *ex ante* or *ex post*. Regardless, the GDPR requires technology providers to address the logic involved in machine learning algorithms. Since these algorithms are not based on logic, but statistical learning, the GDPR introduces a *highly significant* constraint on the future development of artificial intelligence. In fact, given the opacity of machine learning, the key question is not whether the right to "meaningful information about the logic" pertains to systems or their individual decisions, but whether the requirement can be fulfilled at all. In our opinion, we

can answer this in the affirmative, but only if we agree to interpret the constraint as a duty for technology providers to facilitate *rational argumentation* about automated decision-making, aimed at giving substance to the subjects' right to *challenge* those decisions under Article 22.

From a legal point of view, this interpretation – linking the right to information with a right to challenge decisions – is strongly suggested by the language of Article 12-15. In all these provisions, the right to information is conditional on "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4)". Furthermore, the GDPR states only that the right to meaningful information about the logic involved is triggered in "at least" those cases, i.e., the cases referred to in Article 22(1) and (4). This highlights not only an explicit intention of providing a dynamic standard open to interpretation (via the phrase "at least"), it also establishes a tight connection with Article 22.

To convey the significance of the link between the right to information and the right to challenge automated decisions, we quote Article 22 in full below:

(1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

(2) Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

(c) is based on the data subject's explicit consent.

(3) In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

(4) Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(2) 1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

We see that Article 22(1) establishes a limited right *not to be subject* to a "decision based solely on automated processing". However, this right is not without exception, as clarified in Article 22(2). Furthermore, a crucial rule pertaining to this exception is provided in Article 22(3): when the user is subjected to an automated decision that he has not explicitly consented to, he is entitled to "express his or her point of view and to contest the decision". Importantly, this provision pertains to those "cases referred to in Article 22(1)" where the expection in Article 22(2) (a) or (b) is also triggered. Obviously, a right not to be subjected to automated decision-making makes the right to contest such decisions redundant. Hence, in those cases under Article 22(1) where none of the exceptions in Article 22(2) apply, the right to reject the decision as an *unlawful*

violation of Article 22(1) overshadows the right to explanation one might have under Articles 13-15.

It follows that the right to "meaningful information" is particularly relevant for the cases referred to in Article 22(3), which form a proper subset of cases referred to in Article 22(1). Nothing whatsoever can be inferred from the omission of an explicit reference to Article 22(3) in Articles 12-15. In this regard, the interpretation argued for in [11] is fallacious. It is clear from the wording of Article 22 that Article 22(3) pertains to cases addressed in Article 22(1), bringing decision that data subjects are entitled to contest under the scope of the right to obtain meaningful information about the logic involved. Indeed, these are the cases when the right really matters; to contest an automated decision, the data subject *depends on* this kind of information. Hence, instead of regarding those cases as falling outside the scope of Articles 12-15, as [11] does, we conclude that they make up the *core* cases targeted by this provision. As such, the interpretation of what constitutes "meaningful information" must also reflect that the purpose is to provide data subjects with a basis on which to contest automated decisions.

Specifically, in view of the reference made to Article 22(1), we believe the right to meaningful information can be fulfilled by providing users with an adequate basis on which to contest decisions that they have not explicitly consented to, but must nevertheless accept in view of the exceptions in Article 22(2) (a) or (b). Effectively, we believe the right to meaningful information should be interpreted as a constituent part of a more general *right to argue* against automated decisions. In the following sections, we discuss some consequences of this view, related to the interpretation of the "meaningful information" requirement.

Hence, it is our contention that John Doe's new rights entail that he is entitled to an explanation of the decision which is both comprehensible and provides grounds for contesting the decision, when he wishes to do so. This has some implications for how frameworks for automated decision making should be designed and implemented, as we discuss in the following sections. It is natural to develop GDPR-related system constraints, concerning meaning and refutation, on the basis of well-established theories of reason and argumentation, respectively.

3 Theories of mind and mindful theories of machines

When humans *comprehend* the behavior of other humans, they are said to have a theory of mind. The notion of comprehension underpinning this concept is not based on formal logic or epistemology, but on an intuitive ability to recognise mental states that motivate human behaviour. In relation to machine learning, we cannot rely on any similar intuitive understanding of the computational states that motivate certain automated decisions. The problem of providing "meaningful information" about machine behaviour should not be addressed by anthropomorphic explanations that attributed human-like mental states to machines.

However, the deeper question of how intentional behaviour can be understood at all seems to arise in much the same way regardless of whether we address humans, animals, or machines. This is evident when considering the philosophy of mind, which is dominated by theories that draw extensively on analogies between human reasoning and various computational systems. At first sight, this would seem to offer a path towards a methodology for explanation of behaviour that can also satisfy the "meaningful information" requirement of the GDPR. However, as we will argue in the following, our current best theories from the philosophy of mind fail to provide explanation strategies

that are conducive to rational argumentation about decision-making. Importantly, however, these theories highlight certain important features of human reasoning that should influence our assessment of what counts as *meaningful* information to a human that seeks to understand a machine.

Theories from the philosophy of mind attempt to give an explanations of how the mind works. We will discuss two prominent theories here because, in many central regards, they are complementary to each other. Furthermore, they both highlight important challenges that must be overcome to facilitate rational argumentation about automated decisions based on machine learning.

The representational theory of mind

The representational theory [1] posits that we rely on an organized mental model of the world when we relate to it. It puts forth two key ideas about understanding, namely *productivity* and *systematicity*:

Productive Human thought is productive in the sense that there is no principal limit to the depth of our understanding. I am the son of my father, the grandson of my grandfather, the great grandson of my great grandfather, and so on. There seems to be no limit to the number of generations we extend this to, while remaining coherent.[1]

Systematicity There are certain relations between thoughts we *do* have, and toughts we are *able* to have. If we believe the computer is to the left of the cup, we are able to imagine that the cup could have been to the left of the computer. That is, there are certain persistent permutations which seems permissible regardless of whether or not they have ever been encountered.

The representational theory is well-aligned with so-called "classical" notions of thought, emphasising rules of inference and the structure of explanations. However, it has received substantial criticism, particularly in terms of its ability to provide a *causal* explanation of behavior. There has been only limited success in esatblishing credible computer systems that actually *behave similarly* to humans on the basis of this theory. More importantly, while a representational account of symbolic AI systems can be provided in principle, a representational account of machine learning is highly implausible. If information about machine learning algorithms can only be regarded as meaningful when it gives rise to human understanding that satisfies productivity and systematicity, it seems doubtful that machine learning techniques can be considered compatible with the GDPR at all.

Furthermore, even though we are able to replicate, say, systematicity in rule-based systems which are able to say that certain predicates are symmetric (or the like), the implementation of these systems bares little semblance to physical brains, much less neural networks. Hence, we conclude that the representational theory has not yielded a theory of how the brain *functions as* a machine, nor a theory of how the decisions made by a complex machine learning algorithm could be said to have been "understood" by a human under a productivity or systematicity constraint.

[1] If you are concerned that this seems to imply finitude, consider the fact that my great^n grandfather *might* have chased a squirrel which *might* have been searching for nuts, which *might* have fallen on the first day of fall, which *might* have coincided with... any arbitrary event.

Connectionism

In contrast to the representational theory of mind, connectionism stipulates that the mind is essentially like a neural network. According to this theory, mental concepts are represented as distributed patterns of activity over a network, resulting in a theory of sub-symbolic and irreducibly complex mental states. This dispenses with the classical view that the mind is a compositional system where concepts form as combinations of semantic atoms, an idea that features prominently in the representational theory. Still, the theory remains anchored in a computational understanding of mental processes, suggesting again the possibility that our "understanding" of computer systems can be aided by a natural correspondence between how such systems operate and how humans think.

However, there is a major obstacle to this mode of understanding machine learning systems: there is no guarantee that we have any key concepts in common. While humans and machine learning systems might be said to share aspects of the mechanism by which connectionist mental states are realised, this does not entail that the mental states themselves are commensurable. To illustrate this, consider translating a sentence from one language into another. While well-educated adults might be said to rely on representational modes of understanding (grammar, dictionaries etc.), children proficient in both languages might be hypothesised to entertain connectionist representations of the concepts involved, mapping to both languages with equal ease. However, in the case of children, we are hard pressed to explain exactly how they do it [2].

Furthermore, connectionism provides no reason whatsoever to think that the manner in which children translate between human languages bears any similarity to how Google Translate does the same using machine learning techniques [12]. Regardless of the truth of connectionism, the intermediate languages that deep learning algorithms rely on for the purposes of translation might well be incomprehensible to humans [4]. Indeed, the fact that Google Translate has arrived at translation techniques that are hard to understand is arguably the reason why it now performs better than rule-based systems relying exclusively on representational linguistics.

More generally, it seems that the promise and potential of machine learning might be most signficant when it helps us in making predictions and decisions about chaotic systems that humans seem incapable of comprehending analytically [9, 8]. These are instances where neither a representational nor a connectionist account of human understanding can provide pointers to "meaningful information" about the logic involved, since the solution found by the machine is – to the best of our knowledge – inherently meaningless to humans. In these cases, it seems to us that the best we can hope for is *incomplete* and *imprecise* information that simplifies and abstracts away from the details of how the machine reaches decisions. The question is what it takes for this type of information to count as "meaningful information about the logic involed". This is where we think a focus on argumentation is in order, to differentiate between "good" and "bad" forms of incomplete and imprecise information. Essentially, it seems to us that good forms of incomplete and imprecise information are those forms of information that afford data subjects meaningful fulfilment of their right to challenge decisions under Article 22(3) of the GDPR. In the next section, we sketch how (informal) argumentation theory can support an interpretation whereby "meaningful information about the logic involved" can be taken as a requirement on technology providers to supply high-quality enthymematic arguments that track the decision-making of machine learning algorithms as closely as can *reasonably* be expected given our current level of understanding of such systems.

4 On enthymematic arguments

Assume that there exists an ideal argument A^m which perfectly expresses the reason for p, e.g., a perfectly precise and accurate reason why a given machine learning algorithm produces a certain output in a certain context. In this case, we can suggest any A that also concludes with p as a so-called enthymematic argument (with unstated premises) that *approximates* A^m. We can do this even if A^m is not understandable or too complex to ever state in full. Staggering complexity is indeed the typical situation in human argumentation, especially when discussing complex matters like politics, religion or why we "like" a given Facebook post. In these cases, the arguments we use to explain our position provide a necessarily imprecise and incomplete description of why we hold certain beliefs, such as a belief in God, the ideals of social democracy, or the aesthetic appeal of one of our friend's wedding photos. While we labour under the belief that there *is* a perfect A^m explaining our position, what we require of A is that it is both understandable and sufficiently precise. In real-world argumentation, this is what "meaningful information about the logic involved" often amounts to.

To refine what we mean by "sufficiently precise", we rely on perceived conformity to the fragment of A^m that is observable or known. So, for instance, if a belief in fairness guides our belief in either God or social democracy, we are not expected to argue for those positions on the basis of self-interest. Of course, fairness and self-interest might be compatible under some theories, but the contradiction between these sentiments is sufficiently plausible to render it a *prima facie* challenge that can be raised against apparently selfish Christians or seemingly egotistical social democrats. Indeed, part of the reason why humans seem rather preoccupied with the apparent hypocrisy and the general reputation of other arguers is that they depend on such considerations when attemtping to judge whether a given enthymematic argument should be considered a "good" approximation of some ideal argument that sets out a coherent case in full.

In the context of providing meaningful information about automated decision-making, similar considerations can and should be made. The perspective we adopt should be holistic, based on a recognition that the best we can hope for in terms of "meaningful information about the logic involved" are enthymematic arguments about the system that we collectively judge as adequate or not for the purposes of rational argumentation. This shift of perspective takes a significant aspect of what it means to explain automated decisions out of the computer science realm, importing it instead into the realm of the social sciences. This might not be desirable as such, but it is necessary. Even the best human engineers are unable to give a precise causal explanation of the output of a sufficiently complex neural network. Hence, the situation is parallel to that encountered in human argumentation; we are *forced* to abandon the use of either representational (classical) logic or connectionist (causal) descriptions when we wish to communicate reasons for complex behaviors and beliefs.

We are forced to permit vagueness in the description of the computer behavior, which consequently means that we are forced to develop strategies for dealing with this vagueness. In [7, 6], we encounter the notion of the "crater", used to refer to the space of possible interpretations or completions of a given enthymematic argument. Given this terminology, a high-level description of how vagueness should be dealt with is to say that whenever the complete argument A^m is contained in the crater of the actually uttered A, then either:

1. the crater is too large, or

2. the user understands A^m.

When the crater is "too large", we must insist that the actor narrows the space of possible completions and interpretations. This can only be done by dissecting the phenomenon under consideration to produce new enthymematic arguments (with smaller craters) and then recombining them in some way. There is an important balance here between the accuracy we are legally entitled to and the accuracy that actors can afford to provide. Importantly, this trade-off cannot be considered from a vantage point outside the context of the argument and its audience. Considerations pertaining to vagueness and accuracy form critical parts of the dialectic social process by which a group of people exchange arguments about some phenomenon. The idea that sound knowledge is gained at the group level from such processes, despite their adversarial characteristics, is crucial to the so-called *argumentative theory of reason* [5]. This theory claims that reasoning as such evolved to facilitate good arguments conducive to winning debates, not to arrive at the truth. Even so, the theory claims, the collective pursuit of the truth is generally helped by this, as participants adopt a highly critical attitude to claims contradicting their own beliefs (often indicative of confirmation bias at the individual level).

It seems to us that while the philosophy of mind discusses many interesting hypotheses regarding human understanding and computational processes, it is the argumentative theory of reason that provides the best theoretical basis on which to determine the extent of the right to "meaningful information" in the GDPR. Importantly, this theory highlights the importance of argumentation, facilitated by arenas where different actors with conflicting interests are able to engage in meaningful debate. It is our opinion that if the right to meaningful information is interpreted in view of Article 22, it can serve to establish just such an arena, where the asymmetry of power between data subjects and technology providers is offset by the fact that the latter is required by law to facilitate the necessary raw materials that will enable the data subject to argue against the decision made by the system. This is reminiscent of the principle of falsification adopted to great effect in the experimental natural sciences, with one key difference: the testing bed for automated decision-making will not be the laboratory, but the public sphere and – ultimately – the courts. Hopefully, the principle can still serve to facilitate incrimental improvements in our collective understanding of how machine learning works, how it can be improved, and when it is appropriate to apply it. In so far as this objective is reached, moreover, we should hardly hesitate in concluding that we have indeed arrived at a fitting interpretation of what "meaningful information" actually means.

5 Implications for computer science

Modern machine learning is achieving precision and accuracy beyond what many other approaches in AI are capable of. While it would clearly be folly not to utilize these new methods as the precision and accuracy they provide clearly demonstrate their value. Such measures of quality are a necessary condition for their application, but GDPR entails that these measures in themselves are not sufficient in many circumstances. The new EU regulation addresses the cases in which a decision concerning an individual subject is produced automatically. We have argued for why we believe that this new regulation grants the subject a right to meaningful information *in order to* enable her to produce an *informed objection*.

A key difficulty in formulating this consideration as a software constraint, is including a requirement of *meaningfulness*. This issue has not been extraordinarily pressing in

traditional AI. Symbolic, rule based systems are closely related to the representational theory of minds – yielding a transparent relationship between human understanding and computer algorithms. Constructing a decision by composing the rules that constitute such a system can also form an understandable explanation. Even in this case, however, such an argument structure might be simplified in order to enhance comprehension by eliminating obvious or commonly known premises, not including which rules were applied and why, and so on. This is precisely to construct an enthymeme. Notice that even if these rules are strict logical rules in the framework, they may not be admissible when we eliminate premises. Therefore, the simplified explanation will generally also rely on defeasible rules.

Unlike the decisions extracted from the constituent parts of a rule based systems, the decisions from non-symbolic or sub-symbolic algorithms, such as the output of neural networks, must somehow be synthesised in order to comply with the GDPR. Then these decisions may also be connected with the construction of rules which are informative to the subject. One way of achieving such a representation is to find correlations between particular sub-symbolic features and natural predicates which permit us to illustrate conditions which align with the network's output. Regardless, it is up to the system designer to construct a necessarily lossy projection from the neural network's values on to features of natural language which *explain* the output. Regardless of how the projection is constructed, it can not convey the entire decision process. As demonstrated in [9, 8], there are efficient systems that are *inherently* incomprehensible. When we want to provide the required information, we are indeed, producing an enthymeme.

Hence, we need – we are legally obliged – to balance between well-established measures of quality such as precision and accuracy on the one hand, and a meaningful explanation on the other hand. This explanation must necessarily be less complex than the full analysis in many cases. This, we argue, is a consequence of a natural interpretation of the GDPR. Additionally, we believe that there is good reason to consider systems based on defeasible logic as a natural technology to fill this gap.

In the confluence between logic and argumentation, the notion of an enthymeme naturally performs this function in human dialogue. The recent developments in formal argumentation theory seems very well-positioned to help us formalize this challenging balance. When we develop or apply methods for automatic decision making, we need to take this new challenge into account.

6 Conclusion

The clash between the wavefront of enthusiastic innovation in computer technology and conservative forces emphasising the dangers of artificial intelligence are entering an exciting period. Human perception of what is *required* according to a system of rules and what is *achievable* according to a holistic understanding of the world are both put to the test, raising the question of how we should balance effectiveness and fairness when developing the computer systems of the future. This is not a recent problem, but until recently it has been pressing only in fantasy novels and the works of progressive philosophers.

As the GDPR is coming into effect, we are seeing a clash between several competing stances on the very meaning of artificial intelligence; particularly between effective machine learning and understandable rule-based systems. We argue that the distinction is reminiscent of a deep, philosophical divide, but that there is a pragmatic compromise: argumentation theory. We can balance the expressive ability of connectionistic machine

learning approaches which are difficult to relate to on the one hand, and readily comprehensible rule-based approaches generally subscribing to the representation theory of mind.

When the user is able to challenge impenetrable conclusions of machine learning, forcing answers that split explanations up into (potentially non-deductively closed) sets of understandable claims, the user will finally be able to claim the rights she is entitled to; meaningful information about the logic applied in the analysis of her data. As we have argued in this paper, there is a good argument to be made that something along these lines is now a legal entitlement for data subjects under the GDPR.

References

[1] N. Block and J. A. Fodor. What psychological states are not. *Philosophical Review*, 81:159–81, 1972.

[2] S. A. Gelman. Learning from others: children's construction of concepts. *Annual review of psychology*, 60:115–140, 2009.

[3] B. Goodman and S. Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *ArXiv e-prints*, June 2016.

[4] P. Isabelle and G. Foster. Machine translation: Overview. In K. Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, pages 404 – 422. Elsevier, Oxford, second edition edition, 2006.

[5] H. Mercier and D. Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–7, 2011.

[6] F. Paglieri and J. Woods. Enthymematic parsimony. *Synthese*, 178(3):461–501, 2011.

[7] F. Paglieri and J. Woods. Enthymemes: From reconstruction to understanding. *Argumentation*, 25(2):127–139, 2011.

[8] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Phys. Rev. Lett.*, 120:024102, 2018.

[9] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott. Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. *Chaos*, 27(12):121102, 2017.

[10] A. D. Selbst and J. Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 74:233–242, 2017.

[11] S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7:76–99, 2017.

[12] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang,

C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.