Machine Thinking, Fast and Slow

Jean-François Bonnefon, Iyad Rahwan

Version of Record: https://www.sciencedirect.com/science/article/pii/S1364661320302229
Manuscript_c68b903e710c30fd0db38671966cc007

Machine Thinking, Fast and Slow

Jean-François Bonnefon[1]* and Iyad Rahwan[2]*

[1]Toulouse School of Economics (TSM-R), CNRS, Université Toulouse Capitole, Toulouse, France
[2]Center for Humans & Machines, Max-Planck Institute for Human Development, Berlin, Germany

* Correspondence: jean-francois.bonnefon@tse-fr.eu; rahwan@mpib-berlin.mpg.de

Acknowledgments
JFB acknowledges support from the ANR-Labex Institute for Advanced Study in Toulouse, the ANR-3IA Artificial and Natural Intelligence Toulouse Institute, and the grant ANR-17-EURE-0010 Investissements d'Avenir.

Keywords
Artificial intelligence; machine behavior; dual-process; trust; algorithm aversion; machine ethics

Abstract
Machines do not "think fast and slow" in the sense that humans do in dual-process models of cognition. However, the people who create the machines may attempt to emulate or simulate these fast and slow modes of thinking, which will in turn affect the way end users relate to these machines. Here we consider the complex interplay in the way various stakeholders (engineers, user experience designers, regulators, ethicists, and end users) can be inspired, challenged or misled by the analogy between the fast and slow thinking of humans and the Fast and Slow Thinking of machines.

Analogies between human and machine behavior

Machine behaviorists use the methods of the social and behavioral sciences to study intelligent machines as if they were humans or other animals [1]. For example, if machine behaviorists wished to understand the behavior of virtual assistants such as Siri or Alexa, they could run controlled experiments to see how the assistants react to different requests in different contexts. This approach does not assume that Siri and Alexa think the same way as human assistants, or that it would be useful to understand human assistants in order to understand virtual assistants. However, even though machine behaviorists would not draw such a naive analogy between human intelligence and machine intelligence, it would also be a mistake to completely forget that virtual assistants are inspired by human assistants.

First, if the analogy between human and virtual assistants guided the decisions of developers (engineers, user experience designers etc.), then it can be useful to understand how they, the developers, conceive of human assistants to better understand the virtual assistants they created. Second, if the analogy guides the behavior of users, then it can be useful to understand how they, the users, conceive of human assistants, to better understand how these expectations shape their interactions with virtual assistants — which can in turn shape the behavior of the virtual assistants themselves. Thus, folk theories of how humans think can powerfully shape how machines are developed and how these machines are ultimately perceived by users.

One folk theory in particular has permeated popular culture over the last decades: The idea that humans "Think Fast and Slow". Machines do not think fast and slow in the sense that humans do — but the way humans think fast and slow can inspire the people who create intelligent machines, just as they are inspired by other features of the human mind and brain [2]. Furthermore, and more importantly for this article, people have many expectations about the way others Think Fast and Slow, and these expectations can shape their relations with intelligent machines as soon as these machines somehow convey or pretend that they are thinking fast and slow as humans do.

The folk theory of Thinking Fast and Slow

Our purpose does not require a state-of-the-art review of dual-process models, because we are not concerned about the way humans *actually* think fast and slow. What we are concerned with is the folk theory of Fast and Slow Thinking (capitalized, see below) — or maybe the folk *theories* of Fast and Slow Thinking: the theories that AI developers use as a rough analogy for the way their machines process information; the theories that come from exposure to bestsellers such as *Blink [3]*, *Thinking, Fast and Slow [3,4]*, or *Gut Feelings: The Intelligence of the Unconscious [5]*; or simply the theories that come from a lifelong experience with the way other people think and act. Here we adopt the convention of using the terms Slow Thinking and Fast Thinking, capitalized, when we refer to the folk version of the dual-process model. We do not capitalize 'fast' and 'slow' when we talk specifically about decision speed, or when we refer to the dual-process model itself and not its folk version.

We do not know of any systematic work investigating the contents of the folk version of dual-process theory. Note that the question here is not whether people have a *correct* folk theory of dual-process thinking. We have a reasonable consensus about how physics or biology work, and thus it makes sense to ask how accurate is folk physics or folk biology — but we do not have a similar consensus on dual-process thinking, and thus cannot really assess the accuracy of the folk dual-process theory. From a descriptive perspective, it seems a common enough assumption that "Folk psychology [...] may be tracking, obscurely, the same fundamental duality that scientific psychology has identified" or that "The core of dual-process theory is present in the everyday distinction between intuition and reason — the former immediate, quasi-perceptual, sensitive to

subconscious cues and sometimes biased; and the latter slow, effortful, explicit and more cautious" [6]. What is missing though, as far as we know, is an examination of how exactly the everyday distinction between intuition and reason maps onto the formal distinction(s) that psychologists make between intuitive and reflective processes. What we do have, though, is an empirical literature investigating how people respond to cues of intuition or reflection in others, and how these cues inform their perception of others. This is the literature we build on in the rest of this article. There is strong evidence that people use time, effort, and other metacognitive cues linked to dual-process theory to make inferences about humans and machines. In this article, we make the assumption that these inferences reflect a folk theory of the Fast and Slow Thinking of machines and humans — but the contribution we make does not entirely depend on this assumption. Indeed, the issues we review fall under the broader umbrella of the metacognition of machines, that is, the way people project human cognition on machines.

While we defer to later sections the detailed examination of the metacognitive cues that people attend to, and the effects these cues have on their perception of humans and machines, we can briefly summarize here some of the key differences we will discuss between Fast and Slow Thinking. First, Fast and Slow Thinking differ in speed and scope. Slow Thinking is, well, slow, and can only process a small amount of information at a time — whereas Fast Thinking can process large amounts of information without much effort. Second, Slow Thinking is based on explicit rules and transparent reasons — whereas Fast Thinking is implicit and based on opaque heuristics. Third, Slow Thinking plays a corrective role when Fast Thinking is led astray, since Fast Thinking is biased to rely too much on what is usual, normal, or familiar.

Our focus in this article is on mapping how people use the analogy of Slow Thinking and Fast Thinking on the machines they build and use. We distinguish between four groups of stakeholders: engineers; end users; user experience designers; regulators and ethicists. We argue that these groups focus on different facets of the analogy, as we shall explore below (see Figure 1 for a summary).

Engineering: Do Fast and Slow Thinking correspond to different programming techniques?

There is a long history of using the analogy of Fast and Slow Thinking in AI research. Even before the rise of machine learning, AI scientists made a distinction between heuristic programming approaches [7] based on quick-fire rules [8], and symbolic approaches that are based on carefully curated analytical rules [9] subject to formal logic-based reasoning [10,11]. This separation was never clear-cut, though, since even logic-based reasoning often relies on heuristic pruning of combinatorial search spaces, and is frequently applied over a combination of heuristic rules, thus performing a kind of hybrid of Slow and Fast Thinking. Similarly, models

that penalize complexity, such as Lasso regression [12], can be seen as a form of Fast Thinking over simpler models.

AI engineers can also use the nature of the task in order to distinguish heuristic and analytical reasoning. For example, one might consider automated theorem proving as an analytical task, while considering short-term motion planning as a heuristic task, even if one uses heuristics to speed up the former, and symbolic reasoning to conduct the latter.

Nowadays, when AI developers talk about machines Thinking Fast or Slow, they tend to agree that machine learning — and in particular, deep learning [13] — corresponds to Fast Thinking [14,15], based on analogies of speed and scope, rules and reasons, and bias and correction. But often, this categorization fails to distinguish between the process of model learning itself —e.g. using backpropagation to train a convolutional neural network on millions of image-caption pairs to learn a high-dimensional distribution of image features [16] — from the application of the learned model — e.g. feeding a new image to a pre-trained neural network. The learning process may indeed be a very slow process, but it still seems to correspond to Fast Thinking.

Some argue that where machine learning operates on large amounts of data, Slow Thinking would only operate on a sparse network of knowledge [17,18]. Operating on such a sparse network would also help the machine to express its conclusions in terms of explicit causal models with rules and reasons, something that Fast Thinking can struggle with [19,20]. These causal models would also help machines to overcome the problems that Fast Thinking has with out-of-distribution generalizations, that is, correct inferences about situations which are far from usual or familiar [18]. While many researchers agree that machine learning stands for Fast Thinking in the dual-process analogy, there is no comparable agreement about what would stand for Slow Thinking. Some researchers argue that a hybrid approach is required, in which Slow Thinking is handled by a symbolic model, whereas others argue that Slow Thinking can be handled by an appropriately modified machine learning approach [14,15,21,22].

User Experience

Do people prefer machines to Think Fast or Slow?

People do not trust machines equally for all tasks. For example, people trust machines to solve complex numerical problems [23] more than they trust them to recognise good jokes [24], hire good employees [25], or deal with emotions [26]. Generally speaking, it seems that people make a distinction between 'objective' and 'subjective' tasks. They see objective tasks as requiring rules and explicit reasoning; and subjective tasks as requiring intuitions, instinct, and implicit processing. In other words, they see objective tasks as appropriate for Slow Thinking, and subjective tasks as appropriate for Fast Thinking. And as it turns out, they spontaneously trust

machines to perform objective tasks, but not subjective tasks [27]. That is, they trust machines to Think Slow, but not to Think Fast (see Box 1 for a discussion of trust in machines beyond Fast and Slow Thinking).

However, we may be at the exact point in time when this pattern will begin to change. It is indeed possible that until recently, people were simply not accustomed yet to the idea of machines that would Think Fast — that is, machines that would engage in the implicit, holistic processing of a large amount of information, rather than on the sequential, explicit application of a set of rules. What people were accustomed to instead was the idea of machines that would Think Slow, apply logical rules or mathematical formulas in a rigid way, for predictable results. It is no longer rare, though, for discussions of intelligent machines in popular media, to reference so-called black-box algorithms, or to emphasise the impenetrability of deep learning techniques. Accordingly, people are getting more exposure to the idea that some machines may Think Fast. And as it turns out, when people get exposed to the idea that machines may Think Fast, they start trusting them with subjective tasks, instead of just trusting them for objective tasks [27].

Accordingly, we may be at a turning point regarding the tasks people are willing to leave to intelligent machines. People used to have a restricted notion of what these machines were good for: objective, Slow Thinking tasks. But they are now getting used to the idea that machines can think in another way, a way for which they have the perfect analogy: Fast Thinking. People perceive a difference between Fast and Slow Thinking, and they believe that these two modes of thinking are good for different tasks [28]. As a result, when they have to trust a machine, they may not simply ask themselves 'Does the task require Slow Thinking?', but rather 'Can the machine think in whichever way is required by this task?'

Machines that look like they can think both ways, as humans do, would accordingly be more likely to be trusted, since they would seem to be flexible enough to handle both kinds of tasks. This raises at least two questions, though. First, what do people infer about a machine that is showing signs of Thinking Fast or Thinking Slow? And second, should we allow machines to pretend that they are Thinking Fast or Slow, if this pretense could improve their interactions with humans?

What do people infer about machines that Think Fast or Slow?

So far we have suggested that people may react differently to machines they perceive as Thinking Fast or Thinking Slow. If that is true, we need to consider the signals that people may attend to when deciding whether a machine is Thinking Fast or Slow. AI user experience (UX) designers may exploit these signals, in order to influence the way people interact with the AI. This is similar to the way UX designers exploit psychological quirks in their design of social media sites, online shopping sites, or smartphone applications [29].

When humans think, the main signals of Slow Thinking are time and effort. The longer it took to reach a response, and the harder someone seemed to think, the more likely that person engaged in Slow Thinking [30]. People make many inferences based on this perception. They believe that Slow Thinkers are more likely to be correct, at least for difficult questions [28]. They believe that Slow Thinkers have a more complicated moral character than Fast Thinkers — that is, they see a strong link between the actions of Fast Thinkers and their moral character, but they are not so sure about this link for Slow Thinkers [31]. People also brace themselves when someone engages in Slow Thinking before talking, because they see this Slow Thinking as a harbinger of bad news — which are more delicate to phrase than good news [32].

We should not automatically assume that people will make any of these inferences about machines that they perceive as Thinking Slow. For example, when it takes a human a long time to generate a prediction, people trust that prediction more — but they trust the prediction *less* if it took a *machine* a long time to generate it [33]. One possibility is that people judge humans and machines using different benchmarks for decision time. Just as fifteen years is a short life for a human but a long life for a dog, fifteen seconds of thinking may be short for a human, but long for a machine. Accordingly, fifteen seconds may be a signal of Fast Thinking for a human, but a signal of Slow Thinking for a machine — and fifteen *minutes* could be a signal of Slow Thinking for a human, but a signal of malfunction for a machine. Thus, if we want to explore whether people react the same to humans who Think Slow and machines who Think Slow, we will have to carefully calibrate what counts as a signal of Slow Thinking for humans and for machines.

Another signal of Slow Thinking is the explicit correction of one's self-diagnosed mistakes or biases. This signal could play an important role in contexts where machines must make high-stake decisions, for example in the moral domain. Machines can help make decisions with a large impact on the life outcomes of humans — for example, when approving bank loans, deciding whether a defendant will get parole, allocating kidneys, or make split-second choices when driving a car at full speed in a critical situation. In such situations, it is important to ensure that the decisions or recommendations of the machine meet appropriate ethical standards. For that purpose, the distribution of labor is usually that the machine Thinks Fast, and humans Think Slow. Either humans try to consider ethical dilemmas before they occur, in order to ponder them before the machine needs to solve them [34,35], or humans discover ethical issues with the Fast Thinking of the machine, after the fact, and then apply Slow Thinking to correct these issues [36,37]. But what if machines could use some form of Slow Thinking about their own Fast Thinking, in order to diagnose ethical issues with their own behavior, and self-correct to eliminate these issues? How exactly this could be implemented is a matter of debate [15,22], but our focus here is on how it would change the way people perceive intelligent machines that make high-stake decisions.

First, we would need to investigate whether people think that self-diagnosing and self-correcting ethical lapses is a job for Slow Thinking — and if they think this is equally true for humans and for machines. Second, we would have to investigate the inferences people would make about a machine that displays the ability to self-diagnose and self-correct in the moral domain. Consider a machine meant to help a judge to decide which defendant gets parole. Imagine now that after a few months of use, the machine tells the judge that it has diagnosed a racial bias in the recommendations it has made so far, and that it will now try to be more lenient to black defendants. How would the judge react? Would she be more likely or less likely to trust the recommendations of the machine from that point on? It is not at all clear that the admission of past mistakes might increase trust in the machine — insofar as we know that people lose trust in machines that make mistakes faster than they lose trust in humans who make mistakes [38,39]. While this aversion can be overcome if people can tweak the machine themselves [40], we do not know if it can be overcome by a promise by the machine to fix itself.

Law & Ethics: Should we allow machines to pretend they Think Fast and Slow?

We have argued that the trust that people have in machines which perform as advisors or social partners may depend on the kind of Thinking that machines signal to humans. Machines do not really think fast or slow, though, not as humans do. In a sense, people are being deceived, at least by omission, when they react positively to a machine because they have projected on this machine the kind of thinking that humans do.

This means that we have to be careful about where to draw the line when we allow machines to send signals of Fast or Slow Thinking in their interactions with humans. Consider for example the problem of explanation. Many argue that machines would be trusted more if they could articulate a rationale for their predictions or recommendations — instead of, e.g., simply stating these predictions or recommendations based on an opaque neural network computation [19]. Providing an explicit reasoning for a conclusion (instead of framing it as an intuition) is a characteristic of Slow Thinking. In other words, the machine is displaying signals of Slow Thinking in order to be trusted more. Maybe that alone would not count as deception, if the machine provides an explanation that is a genuine and straightforward reflection of the reasons underlying its recommendations. It is perhaps unavoidable that in so doing, it will appear to Think Slow to human eyes — and if this appearance of Slow Thinking makes the machine more convincing, maybe this is just a positive but not explicitly pursued side effect of making the machine more transparent [41].

This is only possible, though, for machines that process information in a way that naturally lends itself to simple explanations — and this is itself an increasingly unlikely occurrence. In most cases, the explanation is only an approximate mapping of what actually went on under the hood. In a sense, the explanation is akin to a data visualization or an infographic. The data visualization

aims at helping viewers to understand the gist of the results, but it does not aim at explaining the intricacies of the statistical modelling that allowed the results to be established. It is constrained by the data, but it is also a persuasion tool: Among the dozens possible ways of visualising the data, one was singled out, the one that seemed the most convincing. The same goes on when a machine produces an explanation of its deep computations. The explanation is the result of a complex trade-off between fidelity, simplicity, and persuasiveness [19]. Consider for example a case where two explanations could be produced. The two explanations have the same fidelity to the actual computations of the machine. However, one of them will not be as convincing to humans, because of a known cognitive bias. In this case, it would seem best for the machine to show users the most convincing of the two explanations [42,43]. After all, why be self-defeating and show humans an explanation that they are irrationally biased against? This is a path, though, that can easily lead to deception. Imagine again that two explanations could be produced. One explanation is slightly more faithful to the Fast Thinking of the machine, but it is also vastly less convincing, still because of a cognitive bias. Should it be eliminated in favor of the explanation that is slightly less faithful but vastly more convincing?

Doing so would be deceptive if users had reasons to expect maximally faithful explanations. It would be less deceptive if users were warned that the machine may at times try to *rationalize* its recommendations, rather than explain them. It would be fascinating to see how much such an admission would change the trust that people have in machines. This change could be small, for psychological and practical reasons. From a psychological perspective, rationalization is a common use of slow thinking in humans [44,45], and it is accordingly possible that people would expect and tolerate some level of rationalization from machines, too, as soon as the machines pretend to Think Fast and Slow. From a practical perspective, we know that people are sometimes willing to be nudged toward a decision that serves their goals, even when the nudge can be construed as manipulative [46]. Accordingly, it is possible that they may allow machines to be slightly manipulative, if such machine behavior is proven beneficial [47].

Concluding remarks

Intelligent machines have become so prevalent, so complex, and so intricately embedded with humans and other machines, that it is no longer possible to predict what they will do based on how they were designed. As a result, scientists have started to use the methods of behavioral science to study intelligent machines from the outside, as if they were humans or other animals. Citizens are engaging in their own version of this paradigm shift. Faced with machines whose capabilities and inclinations have become mysterious, people are recruiting the mental tools they use to assess other humans, and repurposing them for the assessment of machines. In this article, we speculated in particular about the way people relate to machines that display signs of Fast and Slow thinking.

As soon as machines display signs of Thinking Fast or Thinking Slow, people will make inferences about what they can trust the machines with, just as they make these inferences about humans. Accordingly (and see the Outstanding Questions for specific examples), we need to understand the signals of Fast or Slow Thinking that people pick up when they deal with machines; we need to compare the inferences they made about machines that Think Fast or Slow, to the inferences they make about humans; and we need to tackle the difficult ethical concerns that come with the potential for manipulation intrinsic to machines that learn to pretend they are Thinking Fast or Thinking Slow.

Text Box 1: Trust in Machines

This article focuses on how cues of Fast or Slow Thinking can influence people's trust in intelligent machines, but many other factors can affect this trust [48]. For example, people trust intelligent machines more when there is evidence that they perform well on a given task, or when they are routinely used for this task [49]. Anthropomorphism [50] and social behavior [51,52] can help a machine build trust, although these effects can be complex. For example, while machines can learn behaviors that reliably elicit cooperation from humans [53], the mere fact that they are machines seems to penalize the trust they inspire — even when keeping their behavior constant, machines that reveal their true nature elicit lower cooperation than machines who pretend to be human [47]. People trust machines less for decisions which have a moral component [54], although this distrust may be attenuated if they share the ethical values embedded in the machine [55]. If the ethical values embedded in machines are not aligned with the ethical values of users, there is a risk that users will simply opt-out of using these machines, and forfeit the potential collective benefit of their use [56]. Cultural and individual differences can affect the values that people want to see embedded in machines, which makes it important to collect data in large, multicultural samples in order to explore cultural and individual heterogeneity [35,57]. All the factors that affect trust in machines may interact with the way machines display signals of Fast or Slow thinking. Whether the machine Thinks Fast or Slow may affect expectations of performance on different tasks; anthropomorphism may amplify the effect of signals of Fast or Slow Thinking; machines that learn to display signals of Fast or Slow Thinking may be better able to pass as humans; Fast or Slow thinking may be perceived as more or less appropriate or desirable, depending on the moral component involved in the decisions of the machine. Cultural norms may change the preferences of users for machines that Think Fast or Think Slow, and individual differences may have a comparable impact. For example, people with a proclivity to process information reflectively may find it easier to trust machines who Think Slow.

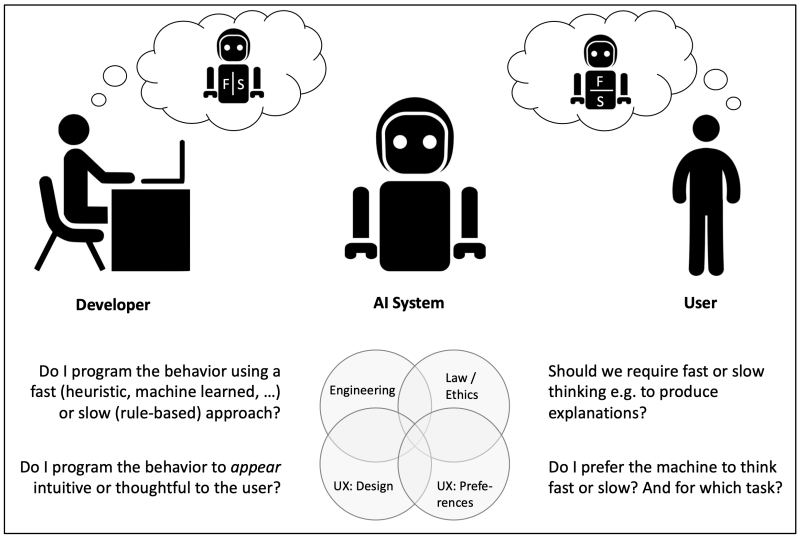
Figure 1. Examples of the use of the 'Fast and Slow' analogy by stakeholders. *The folk theory of mind that people have about machines Thinking Fast or Slow impacts various stakeholders. These stakeholders may not use the analogy in the same way, as symbolized by the different arrangements of F and S in the thought bubbles. On the development side, engineers use the Fast and Slow analogy to select algorithmic solutions, while user experience (UX) designers use the analogy to shape how the machine is perceived by the user. On the user side, regulators might encourage Slow Thinking by machines for the purpose of producing explanations, while end user psychology ultimately determines how signals of Fast and Slow Thinking (e.g. hesitation, self-correction, explanation giving) shape the trust placed into the machine. Yet other examples are considered in the main text.*

References

1 Rahwan, I. *et al.* (2019) Machine behaviour. *Nature* 568, 477–486
2 Dehaene, S. *et al.* (2017) What is consciousness, and could machines have it? *Science* 358, 486–492
3 Gladwell, M. (2007) *Blink: The Power of Thinking Without Thinking*, Hachette UK.
4 Kahneman, D. (2011) *Thinking, Fast and Slow*, Farrar, Straus and Giroux.
5 Gigerenzer, G. (2007) *Gut Feelings: The Intelligence of the Unconscious*, Penguin.
6 Frankish, K. (2010) Dual-process and dual-system theories of reasoning *Philos. Compass* 5, 914–926
7 Cohen, P.R. (1985) *Heuristic Reasoning about Uncertainty: An Artificial Intelligence Approach*, Pitman, Boston and London.
8 Brooks, R.A. (1999) *Cambrian Intelligence: the early history of the New AI*, MIT Press.
9 Jackson, P. (1998) *Introduction to Expert Systems*, Addison-Wesley Longman Publishing.
10 Lloyd, J.W. (2012) *Foundations of Logic Programming*, Springer Science & Business Media.
11 Van Emden, M.H. and Kowalski, R.A. (1976) The Semantics of Predicate Logic as a Programming Language. *J. ACM* 23, 733–742
12 Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* 56, 267—288.
13 LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
14 Chen, D *et al.* (2019) Deep reasoning networks: Thinking fast and slow. *arXiv:1906.00855*
15 Rossi, F. and Loreggia, A. (2019) Preferences and ethical priorities: Thinking fast and slow in AI. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (El Fallah Seghrouchni, A. *et al*., eds), pp. 3-4, International Foundation for Autonomous Agents and Multiagent Systems.
16 Krizhevsky, A. *et al.* (2012) ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (Pereira, F. et al., eds), Curran Associates.
17 Anthony, T. *et al.* (2017) Thinking fast and slow with deep learning and tree search. , in *Advances in Neural Information Processing Systems 30* (Guyon, I. *et al,* eds), pp. 1097-1105, Curran Associates.
18 Bengio, Y. (2019) The consciousness prior. *arXiv:1709.08568*
19 Weld, D.S. and Bansal, G. (2019) The challenge of crafting intelligible intelligence. *Commun. ACM*, 62, 70–79
20 Lage, I. *et al.* (2019) An evaluation of the human-interpretability of explanation. *arXiv:1902.00006*
21 Dubois, D. and Prade, H. (2019) Towards a Reconciliation Between Reasoning and Learning - A Position Paper. *Lecture Notes in Computer Science* 11940, 153–168
22 LaCroix, T. and Bengio, Y. (2019) Learning from Learning Machines: Optimisation, Rules, and Social Norms. *arXiv:2001.00006*
23 Logg, J.M. *et al.* (2019) Algorithm appreciation: People prefer algorithmic to human

judgment. *Organ. Behav. Hum. Dec.* 151, 90–103
24 Yeomans, M. *et al.* (2019) Making sense of recommendations. *J. Behav. Decis. Making* 32, 403–414
25 Diab, D.L. *et al.* (2011) Lay Perceptions of selection decision aids in US and non-US samples. *Int. J. Sel. Assess.* 19, 209–216
26 Waytz, A. and Norton, M.I. (2014) Botsourcing and outsourcing: Robot, British, Chinese, and German workers are for thinking—not feeling—jobs. *Emotion* 14, 434–444
27 Castelo, N. *et al.* (2019) Task-dependent algorithm aversion. *J. Marketing Res.* 56, 809–825
28 Kupor, D.M. *et al.* (2014) Thought calibration. *Soc. Psychol. Pers. Sci.* 5, 263–270
29 Gray, C.M. *et al.* (2018) The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Mandrick, R. L. *et al.*, eds), pp. 1–14, ACM Press
30 Mata, A. and Almeida, T. (2014) Using metacognitive cues to infer others' thinking. *Judgm. Decis. Mak.* 9, 349–359
31 Bonnefon, J. F. (2017) *Reasoning Unbound*, Palgrave McMillan
32 Bonnefon, J. F. *et al.* (2015) Some but not all dispreferred turn markers help to interpret scalar terms in polite contexts. *Think. Reasoning* 21, 230–249
33 Efendić, E. *et al.* (2020) Slow response times undermine trust in algorithmic (but not human) predictions. *Organ. Behav. Hum. Dec.* 157, 103–114
34 Bonnefon, J. F. *et al.* (2016) The social dilemma of autonomous vehicles. *Science* 352, 1573–1576
35 Awad, E. *et al.* (2018) The moral machine experiment. *Nature* 563, 59–64
36 Wong, P.-H. (2019) Democratizing algorithmic fairness. *Philos. Tech.* 33, 225–244
37 Srivastava, M. *et al.* (2019) Mathematical notions vs. human perception of fairness. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Teredesai, A. *et al.*, eds), pp. 2459-2968, ACM Press
38 Dietvorst, B.J. *et al.* (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* 144, 114–126
39 Prahl, A. and Van Swol, L. (2017) Understanding algorithm aversion: When is advice from automation discounted? *J. Forecasting* 36, 691–702
40 Dietvorst, B.J. *et al.* (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Manage. Sci.* 64, 1155–1170
41 Stoyanovich, J. *et al.* (2020) The imperative of interpretable machines. *Nat. Mach. Intell.* 2, 197-199
42 Miller, T. (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267, 1–38
43 Kim, T.W. and Duhachek, A. (2020) Artificial Intelligence and persuasion: A construal-level account. *Psychol. Sci.* 31, 363–380
44 Mercier, H. and Sperber, D. (2017) *The Enigma of Reason*, Harvard University Press.
45 De Neys, W. (2020) Rational rationalization and System 2. *Behav. Brain Sci.* 43, e34
46 Sunstein, C.R. (2017) *Human Agency and Behavioral Economics: Nudging Fast and Slow*, Springer.
47 Ishowo-Oloko, F. *et al.* (2019) Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nat. Mach. Intell.* 1, 517–521
48 Glikson, E. *et al.* (2020) Trust in Artificial Intelligence: Review of empirical research. *Acad. Manag. Ann.* 14, 627–660

49 Kramer, M.F. *et al.* (2018) When do people want AI to make decisions? In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (Furman, J. *et al.,* eds), pp. 204-209, ACM Press
50 Waytz, A. *et al.* (2014) The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* 52, 113–117
51 Traeger, M.L. *et al.* (2020) Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proc. Natl. Acad. Sci. U.S.A.* 117, 6370–6375
52 Rahwan, I. *et al.* (2020) Intelligent machines as social catalysts. *Proc. Natl. Acad. Sci. U.S.A.* 117, 7555–7557
53 Crandall, J.W. *et al.* (2018) Cooperating with machines. *Nat. Commun.* 9, 233
54 Bigman, Y.E. and Gray, K. (2018) People are averse to machines making moral decisions. *Cognition* 181, 21–34
55 Freedman, R. *et al.* (2020) Adapting a kidney exchange algorithm to align with human values. *Artif. Intell.* 283, 103261
56 Bonnefon, J. F., Shariff, A., Rahwan, I. (2020) The moral psychology of AI and the ethical opt-out problem. In *The Ethics of Artificial Intelligence* (Liao, S. M., ed), pp. 109-126, Oxford University Press
57 Awad, E. *et al.* (2020) Crowdsourcing moral machines. *Commun. ACM* 63, 48–55


F
S
F
S
F
S
Developer
AI System
User
Do I program the behavior using a fast (heuristic, machine learned, …) or slow (rule-based) approach?
Do I program the behavior to *appear* intuitive or thoughtful to the user?
Engineering
Law / Ethics
UX: Design
UX: Prefe-rences
Should we require fast or slow thinking e.g. to produce explanations?
Do I prefer the machine to think fast or slow? And for which task?