A Two-tiered View on Acceptance

Joëlle Proust

Institut Jean-Nicod
Fondation Pierre-Gilles de Gennes pour la Science,
Ecole Normale Supérieure
29 Rue d'Ulm,
75005 Paris, France
joelle.proust@ehess.fr

Abstract

Experimental studies in metacognition indicate that a variety of norms are used by humans and some non-human agents to control and monitor their cognitive performances, such as accuracy, comprehensiveness, intelligibility, coherence, relevance, or consensus. This diversity of epistemic norms motivates a revision of the concept of acceptance. First, there are different forms of acceptance, corresponding to the specific epistemic norm(s) that constitute(s) them. Furthermore, acceptances need to include a strategic component, from which the epistemic component is insulated, whose function is to adjust the epistemic output to expected utility. Experimental evidence suggests that this two-tiered analysis of acceptance is empirically adequate. Relevance to AI is briefly discussed.

Acceptance and its Norms

Intelligent agency requires an ability to control and monitor one's cognitive states, e.g. retrieve memories, check one's perceptions or one's utterances. The aim of cognitive control is to acquire cognitively reliable properties, such as retrieving a correct answer. Intelligent agents in realistic settings, however, whether natural or artificial, need to track other epistemic norms beyond accuracy, such as the comprehensiveness of a list, the intelligibility of a text, the coherence of a story, the relevance of a remark, or the consensuality of a claim. Experimental studies in metacognition suggest that such norms are indeed used by human and some non-human agents to control and monitor their own cognitive performance (Goldsmith and Koriat, 2008, Couchman et al. 2010). Furthermore, the particular cognitive task in which performance is being monitored has been shown to dictate which epistemic norm is appropriate to a given context.

The goal of this article is to sketch a theory of acceptance that takes advantage of these studies. Acceptances, in contrast with beliefs, are generally recognized as voluntary (Jeffrey 1956, Stalnaker 1987, Bratman 1999, Lehrer 2000, Velleman 2000, Frankish, 2004). Accepting is an epistemic action, involving deliberation, i.e. various forms of cognitive control and their associated norms. There is no consensus, however, about the norm(s) of acceptances. While for Velleman (2000) accepting is regarding a proposition *P* as true, even though it may not be "really true", Cohen takes acceptance to be "a policy for reasoning, (..) the policy of taking it as a premise that *P*" (1992, 5, 7). For Stalnaker, "sometimes it is reasonable to accept something that one knows or believes to be false". Circumstances where this is reasonable include cases where *P* "may greatly simplify an inquiry", where *P* is "close to the truth", or "as close as one needs to get for the purposes at hand". Granting that accepted propositions are subject to contextual variation in their sensitivity to evidence and truth, they cannot be freely agglomerated in a coherence-preserving way, in contrast with beliefs (Stalnaker 1987). Finally, Bratman (1999) claims that acceptances conjoin epistemic and practical goals.

These features of acceptance, however, fail to offer an intelligible and coherent picture of the epistemic action of accepting, and of its role in practical reasoning and decision-making. First, it is left unclear how a context of acceptance is to be construed in a way that justifies applying fluctuating epistemic standards. Second, how can one possibly conjoin an *epistemic requirement*, which is essentially passively recognized and applied, and *utility considerations,* which require an active decision from the agent as to what ought to be accepted in the circumstances?

The Context Relevant to Accepting *P*

Why is accepting *contextual*, in a way that judging is not? Merely saying that acceptances, "being tied to action" (Bratman, 1999), are sensitive to practical reasoning, is not a viable explanation: other mental actions, such as judgments, also tied to action, do not adjust their contents to considerations of practical reasoning. Saying that they are context-dependent because coherence, consistency, and relevance apply within the confines of the existing plan, rather to a theoretical domain, does not explain how epistemic correctness survives instrumental adequacy.

Our first proposal consists in the following claim: Utility may dictate the norm of acceptance relevant to a given context of action, without dictating the output of the corresponding acceptance. As said above, accepting *P* can be driven, among other things, by a goal of comprehensiveness, accuracy, intelligibility, consensus or coherence. For example, you may accept that the shopping list you just reconstructed from memory is comprehensive (all the items previously listed are included), but not that it is accurate (new items are also mistakenly listed). On the proposed view, an acceptance is always indexed to its specific norm: a proposition is never merely accepted, it is rather accepted$_{at}$ or accepted$_{ct}$ etc. (where *at* is short for: accurate truth, and *ct* for comprehensive truth).

Although the selection of a particular epistemic goal responds to the practical features of one's plan, there is no compromise between epistemic and instrumental norms concerning the *content* of acceptances. Agents' epistemic confidence in accepting$_n$ *P* (accepting *P* under norm *n*) is not influenced by the cost or benefit associated with being wrong or right. Thus we don't need to endorse the view that an epistemic acceptance of *P* is yielding to utility considerations, as Bratman suggests.

This proposal offers a natural way out of a puzzle, called the preface paradox, which is raised by traditional views about acceptance: A writer may rationally accept that each statement in his book is true, while at the same time rationally accepting that his book contains at least one error (Makinson 1965). This puzzle is dissolved once it is realized that the author's epistemic goal is one of offering an ideally comprehensive presentation of his subject matter: it will thus not be contradictory for him to accept$_{ct}$ all the sentences in her book, while accepting$_{pl}$ (accepting as plausible or likely) that one of them is false. Hence, a mental act of acceptance$_{ct}$ does not allow aggregation of truth, because its aim is exhaustive (include all the relevant truths) rather than accurate truth (include only truths). Similarly, in the lottery puzzle, an agent accepts$_{at}$ that there is one winning ticket in the one thousand tickets actually sold. It is rational for her, however, not to accept$_{pl}$ that the single ticket she is disposed to buy is the winning one.

From Epistemic to Strategic Acceptance

The *output* of an epistemic acceptance so construed needs, however, to be adjusted to the final ends of the agent's plan. The decision to act on one's epistemic acceptance, i.e., strategic acceptance, constitutes a second, distinct step in accepting *P*. On our view, utility does not just influence the selection of certain epistemic norms of acceptance. It also influences decision in a way that may depart greatly from the cognitive output of epistemic acceptance.

The first argument in favor of this two-step view of acceptance is conceptual. The existence of an autonomous level of epistemic acceptance enables agents to have a stable epistemic map that is independent from local and unstable instrumental considerations. Thus, it is functionally adaptive to prevent the contents of epistemic evaluation from being affected by utility and risk. Second, empirical evidence shows that agents are indeed able to adjust their cognitive control both as a function of their confidence in accepting P, and of the strategic importance of the decision to be finally taken. In situations where agents are forced to conduct a cognitive task, strategic acceptance is ruled out: agents merely express their epistemic acceptance. In contrast, when agents can freely consider how to plan their action, given its stakes, they can refrain from acting on the unique basis of their epistemic acceptance. A decision mechanism is used to compare the probability for their acceptance being correct and a preset response criterion probability, based on the implicit or explicit payoffs for this particular decision. Here agents are allowed to strategically withhold or volunteer an answer according to their personal control policy (risk-aversive or risk-seeking), associated with the cost or benefit of being respectively wrong or right (Koriat and Goldsmith, 1996). A third reason in favor of our two-tiered view is that strategic acceptance can be impaired in patients with schizophrenia, while epistemic acceptance is not (Koren et al. 2006): this suggests, again, that epistemic and strategic acceptances are cognitively distinct steps.

Discussion

The two-step theory sketched above accounts nicely for the cases of acceptances discussed in the literature. Judging *P* true flat-out is an accepting under a stringent norm of accurate truth, while "judging *P* likely" is an accepting under a norm of plausibility, conducted on the background of previous probabilistic beliefs regarding *P*. Adopting *P* as a matter of policy divides into accepting a set of premises to be used in collective reasoning under a norm of consensus, and accepting it under a norm of coherence, (as in judgments by contradiction, legal reasoning, etc.). Assuming, imagining, supposing do not automatically qualify as acceptances. Only their controlled epistemic forms do, in which case they can be identified as forms of premising.

This theory predicts that errors in acceptances can be either strategic or epistemic. Strategic errors occur when selecting an epistemic norm inappropriate to a context, (for example, trying to reconstruct a shopping list accurately, when comprehensiveness is sufficient), or when incorrectly setting the decision criterion given the stakes (taking an epistemic decision to be non-important when it objectively is, and reciprocally). Epistemic errors, in contrast, can occur both in applying a given norm to its selected material, (for example, seeming to remember that *P* when one does not) or in forming an incorrect judgment of confidence about one's epistemic performance (for example, being confident in having correctly remembered that *P* when one actually failed to do so). Appropriate confidence judgments have an extremely important role as they help filter out a large proportion of first-order epistemic mistakes (illusory remembering, poor perceivings, incoherent or irrelevant reasonings etc.).

Is our two-tiered theory relevant to AI research? Although the author has no personal competence in this domain, it appears to be clearly the case. The two-tiered theory of acceptance is inspired by empirical research on epistemic self-evaluation, and by requirements on epistemic reliability. For these reasons, it should help epistemic agency to be modeled in a more realistic way, and conferred to artificial systems with the suitable cognitive complexity. Indeed an artificial agent, for example a social robot, should be able to monitor the epistemic responses of others as well as its own not only for their accuracy, but also for their comprehensiveness, their coherence, and the consensus in a group. Monitoring relevance in speech may, at present, be a trickier issue. Even artificial agents with no linguistic ability, as far as they need to evaluate whether they can solve a given problem, need to have various forms of acceptance available (e.g.: do they have the resources to create a path to a solution? Should they list, rather, all the solutions already used in similar contexts? How plausible is it that one of these solutions is applicable here and now? Epistemic planning depends on such acceptances being alternative options). Furthermore, it is crucial for artificial system designers to clearly distinguish an epistemic evaluation, which only depends on internal assessment of what is known or accessible, and a utility evaluation, which varies with the importance of the task.

Our two-tiered theory of acceptance raises potential objections that will be briefly examined.

Acceptance Does Not Form a Natural Kind?

It might be objected that, if acceptance can be governed by epistemic norms as disparate as intelligibility, coherence, consensus and accuracy, it should not be treated as a natural kind. In other words, there is no feature common to the various forms of acceptance, and for that very reason, the concept of acceptance should be relinquished. To address this objection, one needs to emphasize that normative diversity in acceptances is observed in metacognitive studies: agents, according to circumstances, opt for accuracy or comprehensiveness, or use fluency as a quick, although loose way, of assessing truthfulness (Reber and Schwarz 1999). What makes accepting a unitary mental action is its particular function: that of adjusting to various standards of utility the cognitive activity associated with planning and acting on the world. This adjustment requires both selecting the most promising epistemic goal, and suppressing those acceptances that do not meet the decision criterion relevant to the action considered.

Sophistication implausible?

A second objection might find it implausible that ordinary agents have the required sophistication to manage acceptances as described, by selecting the kind of epistemic acceptance that is most profitable given a context of planning, by keeping track of the implicit or explicit payoffs for a particular option, and by setting on their basis their response criterion.

It must be acknowledged that agents do not have in general the conceptual resources that would allow them to identify the epistemic norm relevant to a context. Acceptances, however, can be performed under a given norm without this norm being represented explicitly. Agents learn to associate implicitly a given norm with a given cognitive task and context: their know-how is revealed in their practical ability to monitor their acceptances along the chosen normative dimension (Perfect and Schwartz, 2002).

Concerning decision making, robust evidence indicates that the ability to re-experience an emotion from the recall of an appropriate emotional event is crucial in integrating the various values involved in an option (Gibbard 1990, Bechara, Damasio and Damasio 2000). Agents are guided in their strategic acceptance by dedicated emotions (with their associated somatic markers), just as they are guided in their epistemic acceptance by dedicated noetic feelings. (Koriat 2000, Hookway 2003, Proust 2007). The probabilist information about priors, on the other hand, seems to be automatically collected at a subpersonal level (Fahlman, Hinton and Sejnowski 1983).

Value Pluralism and Epistemological Relativism

Finally, epistemologists might observe that such a variety of epistemic standards pave the way for epistemic value pluralism, i.e., the denial that truth is the only valuable goal to pursue. Our variety of epistemic acceptings should indeed be welcome by epistemic value pluralists, who claim that coherence, or comprehensiveness, are epistemic goods for their own sake (Kvanvig 2005). It is open to epistemic value monists, however, to interpret these various acceptances as instrumental steps toward acceptance$_{at}$, i.e. as epistemic desiderata (Alston 2005). The present project, however, is not the epistemological study of what constitutes success in inquiry. It rather aims to explore the multiplicity of acceptances open to natural or artificial agents, given the informational needs that arise in connection with their final ends across multiple contexts.

The proposed two-tiered theory of acceptance does not invite a relativistic view of epistemic norms, but rather combats it: Even though agents can accept propositions under various norms, there are rational constraints on norm selection. For example, it may be rational to look for accurate retrieval when making a medical decision, while looking for comprehensive retrieval when shopping. Once a norm has been identified as instrumentally justified, acceptance can only be successful if it is conducted under the epistemic requirements prescribed by its norm. Thus agents, whether natural or artificial, can build, through acceptances, a stable epistemic representation of the facts relevant to their action that is not contaminated by the strategic importance of the decisions to be taken on its basis.

Conclusion

Given the limited cognitive resources available to agents at any given time, it is rational for them to focus on the epistemic goals that will maximize their epistemic potential both in terms of correctness and utility. Our two-tiered theory of acceptance accounts for this consequence of bounded rationality. Our future work aims to clarify the informational basis on which the various epistemic norms operate. We also aim to study norm sensitivity in agents from different cultures, measured by their confidence judgments about particular tasks and performances. Finally, we will study how potential conflicts for a given acceptance between epistemic norms can be generated, and how they are overcome.

Acknowledgments

Thanks to Richard Carter for his linguistic revision and critical discussion of a prior version, and Anne Coubray, Martin Fortier and Mouhamadou el Hady Ba for useful feedback. This research has ben made possible by an ERC advanced grant, "Dividnorm" project # 269616, conducted at the Fondation Pierre-Gilles de Gennes pour la Recherche, Ecole Normale Supérieure, Paris.

References

Alston, W. 2005. *Beyond "Justification": Dimensions of Epistemic Evaluation*, Ithaca, NJ: Cornell University Press.

Bechara, A. Damasio, H. and Damasio, A.R. 2000. Emotion, Decision Making and the orbitofrontal cortex. *Cerebral Cortex,* 10: 295-307.

Bratman, M.E. 1999. *Faces of intention*. Cambridge: Cambridge University Press.

Cohen, L. J. 1992. *An essay on belief and acceptance*. Oxford: Clarendon Press.

Couchman, J.J., Coutinho, M.V.C., Beran, M.J., Smith, J.D. 2010. Beyond stimulus cues and reinforcement signals: A new approach to animal metacognition. *Journal of Comparative psychology*, 124: 356-368.

Fahlman, S.E., Hinton, G.E. and Sejnowski, T.J. 1983. Massively parallel architectures for A.I.: Netl, Thistle, and Boltzmann machines. *Proceedings of the National Conference on Artificial Intelligence*, Washington DC: 109-113.

Frankish, K. 2004. *Mind and supermind*. Cambridge: Cambridge University Press.

Gibbard, A. 1990. *Wise Choices, Apt Feelings*. Oxford: Clarendon Press.

Hookway, C. 2003. Affective States and Epistemic Immediacy. *Metaphilosophy* 34: 78-96.

Jeffrey, R. C. 1956. Valuation and acceptance of scientific hypotheses. *Philosophy of Science,* 23, 3: 237-246.

Koren, D., Seidmann, L.J., Goldsmith, M. and Harvey P.D. 2006. Real-World Cognitive—and Metacognitive—Dysfunction in Schizophrenia: A New Approach for Measuring and Remediating More ''Right Stuff'', *Schizophrenia Bulletin,* 32, 2: 310-326.

Koriat, A. 2000. The Feeling of Knowing: some metatheoretical Implications for Consciousness and Control. *Consciousness and Cognition*, 9: 149-171.

Koriat, A. and Goldsmith, M. 1996. Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review,* 103, 3: 490-517.

Goldsmith, M. and Koriat, A. 2008. The strategic regulation of memory accuracy and informativeness. *The Psychology of Learning and Motivation*, 48, 1-60.

Kvanvig, J. 2005. Truth is not the primary epistemic goal, in Steups, M. and Sosa, E. eds. *Contemporary Debates in Epistemology*, Oxford: Blackwell, 285-296.

Lehrer, K. 2000. Discursive Knowledge. *Philosophy and Phenomenological Research*, 60, 3: 637-653.

Makinson, D. C. 1965. *Paradox of the Preface*, *Analysis,* 25: 205-207.

Perfect, T.J. and Schwartz, B.L. 2002. *Applied Metacognition*. Cambridge: Cambridge University Press.

Proust, J. 2007. Metacognition and metarepresentation : is a self-directed theory of mind a precondition for metacognition ? *Synthese*, 2: 271-295.

Proust, J. forthcoming. Mental acts as natural kinds, in: T. Vierkant, A. Clark, J. Kieverstein Eds. *Decomposing the Will*. Oxford: Oxford University Press.

Reber, R. and Schwarz, N. 1999. Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, 8: 338-342.

Stalnaker, R. 1987. *Inquiry*. Cambridge: MIT Press.

Velleman, J.D. 2000. *The possibility of practical reason*. Ann Harbor: The University of Michigan Library.