Actively open-minded thinking in politics

Jonathan Baron*

Abstract

The concept of actively open-minded thinking (AOT) provides standards for evaluation of thinking, which apply both to our own thinking and to the thinking of others. AOT is important for good citizenship for three reasons: it provides a prescription for individual thinking about political decisions; it serves as a social norm (when others agree); and, perhaps most importantly, it provides a standard for knowing which sources to trust, including politicians and pundits. I provide a current account of AOT as a general prescriptive theory that defines a standard or norm for all thinking, with emphasis on its role in the judgment of the thinking of others, and in maintaining appropriate confidence. I also contrast AOT with other standards. AOT does not assume that more thinking is always better, and it implies that low confidence in the results of thinking is often warranted and beneficial. I discuss the measurement of AOT and its relation to politics. Finally, I report two preliminary studies of AOT in judgments of others thoughts, and the role of confidence.

*Department of Psychology, University of Pennsylvania, 3720 Walnut St., Philadelphia, PA, 19104. Email: baron@upenn.edu.

I thank Gord Pennycook for very helpful comments.

1 Introduction

The problems of politics are worldwide: In many poor countries, government simply does not function; in other countries, the function of government is hampered by widespread corruption, "crony capitalism", and the social norms that support these practices; even in some of the richest democracies, voters accept unsupportable theories about the nature of their problems, leading to the adoption of policies that oppose the well-being of the citizens who supported them; isolationist policies not only hurt the countries that adopt them but also lead to a vicious circle of retaliation in which all are denied the benefits of international cooperation; and world institutions themselves are weak and fragmented, preventing the sort of coordination and log-rolling across issues that could enable reforms. This happens at a time when the human population, and its use of resources, have expanded to the point where their effects on the environment seriously threaten further increases in the world's standard of living (Dasgupta, Erlich, et al., 2013). We are now depending on scientific, technological, and administrative advances not yet made, or fully known, in order to provide sufficient food, water, and energy even for the population we have (Godfray, Beddington, et al., 2010).

The means for solving most of these problems lie in the hands of individual citizens, through their participation in democratic politics, as well as the leaders they elect. Exhortation of individual self-sacrifice cannot do much more than it already has. It is not too much to expect people to recycle their trash (even in the absence of penalties for not doing it) or to reduce their water use during a drought. But the level of sacrifice required to prevent continued unsustainable draining of aquifers is much greater. It requires a combination of government coercion and government support of alternative water sources.

I have suggested (Baron, 2018b) that, in order for a democracy to function well (both for its own citizens and outsiders), its citizens need to endorse three (somewhat synergistic) social norms, which I called cosmopolitanism, anti-moralism, and actively open-minded thinking (AOT). The first two involve specific content, so they are not of concern here, except that they may both be facilitated by AOT.[1] The third is the main topic I shall address here. Note that AOT has three functions. It is a

[1]Cosmopolitanism is a continuum of breadth of political concern, one end of which is a concern for all humanity (or all sentient life) now and in the future. Singer (1982) has argued that it arises, in history, in part from critical reflection on more parochial forms of thinking such as racism, sexism and nationalism. Anti-moralism is opposition to the imposition of moral principles on others when the principles themselves depend on commitments of faith that cannot be defended to those who are affected and those who have not made similar commitments. Such opposition to moralism is expected as a direct effect of the role of AOT in judgment that I describe here.

norm (a standard) for evaluation of thinking, a set of individual dispositions to think in accord with the norm, and a standard for evaluation of other people's thinking, particularly the trustworthiness of sources that claim authority. As a norm, it can also be a "social norm" when it is socially supported. Past literature on AOT has mostly concerned AOT as a norm and a set of dispositions. In this article, I try to focus more on the last property of AOT, its use for evaluation of other people's thinking. Such evaluation is necessary because citizens cannot think through all the relevant issues themselves, at least not anymore (Baron, 1993). We must rely on others: journalists, institutional and religious leaders, scholars and professionals, ordinary people, and politicians themselves.

2 AOT and its measurement

The idea of AOT has many historical precedents in philosophy and psychology. My own attempt to state this idea more precisely began with *Rationality and intelligence* (1985) and was developed further in several editions of *Thinking and deciding* (2008) and elsewhere. It is about thinking, which is what we do when we are unsure what to do or believe.

2.1 Rational thinking

The basis of the initial statement was a theory of rational thinking. It starts with a simple framework for the description of all thinking, regardless of topic, and then applies standard normative models from decision theory to major parameters of thinking within that framework. These parameters are potentially under the control of the thinker; they make no reference to things that the thinker cannot yet know (such as "the right answer"). The result is a prescriptive model, AOT, designed so that, in principle, it is meaningful for someone to try to follow its prescriptions. It is not about the criteria of success and failure (as in the maxim "buy low, sell high"), but, rather, about the processes under the thinker's control. The model does not include any additional domain-specific knowledge. Of course, thinking is more effective when such knowledge is available.

The proposed framework analyzes thinking into *search* for various objects (described shortly) and *inference* from the results of the search. These two steps are often interleaved through several episodes of search and inference. The objects consist of *possibilities* (candidate answers to the question that inspired the thinking, each with a *value* property), *evidence* (objects that bear on the

value of the possibilities), and *goals* (criteria that determine how each piece of evidence affects the value of each possibility). Thus, the value of a possibility is a specific function of the evidence and the goals.

I also assume a representation of *confidence*, which is a measure of the the relative success (in some sense) of the thinking done so far. Confidence is defined differently for beliefs (what to believe) and decisions (what to do), although in both cases "maximal confidence" implies that further thinking is useless. For belief, confidence is "degree of belief", often represented by probability.

For decisions, confidence should depend on the current option (the one that would be chosen if no further thinking is done) and the expected superiority of other options. In the world of expected utility, we might imagine that "expected superiority" should depend on $\sum_i P_i \, p(U_i > U_o) \, (U_i - U_0)$ where P_i is the probability that option i will be chosen as a result of further thinking, U_i is the expected utility of option i, U_0 is the expected utility of the current option, and $p(U_i > U_o)$ is the probability, for each possible option i, that it's expected utility will turn out to be greater than U_0 when all available evidence is at hand. In other words, this is the expected improvement from further thinking. (This is not a proposal that possible superiority should ever be calculation but just an attempt to clarify what confidence might mean for decisions.) If confidence is 100% when the possible superiority is zero, then we might define 0% confidence as the expected utility of the option that would have been chosen with no thinking at all.

Note that decision confidence should be quite high when there are many options that might do better than the current one but the degree of superiority of each one is small.[2] Confidence should also be higher when the current option appears to be much better than any alternatives. Confidence should be lower when a little more search for evidence might show that some other option is much better than the current one.

Search may be directed in various ways, but the relevant general parameters of interest are its amount and its direction, with direction defined as for or against a currently favored possibility.[3] The direction parameter may also be applied to inference, so that the effect of evidence on the value of possibilities is biased toward or against the strongest possibility. But the normative model specifies

[2]von Winterfeldt and Edwards (1986) call this situation a "flat maximum."

[3]The direction parameter for search is not always under the thinker's control. When a doctor orders a test, she does not know in advance whether the result will support her current diagnosis or not.

that it should be neutral.[4] Search is "fair" when it optimally distinguishes possibilities (possible answers to the question at hand).

The optimal amount of search is the largest amount at which the expected benefit of additional search, in terms of its effects on the final conclusion, is greater than its cost.[5] Search itself has costs in time and in other resources that could be used elsewhere. The cost of search depends on factors such as time pressure (high cost of delay), whether the search is enjoyable or painful in its own right, and whether it involves the use of external sources. The benefit of additional search is the expected utility of stopping thought after additional search minus the expected utility of choosing without additional search (i.e., stopping the search and making do with what has been found). In a decision, if the same option would be chosen before and after additional search, then the search has no benefit for the choice alone. Thus, it is usually wasteful to look for reasons why a favored option is best, if one is going to choose that option anyway. And, for trivial decisions, thinking may have very little value and the optimal amount of it may be zero.

The expected utility of additional search is higher (other things being equal) when confidence in the thinking done so far is lower, since there is more room to increase, and when the expected (justified) increase in confidence from additional search is higher. Thus "confidence" has opposite effects on search depending on which sort of confidence is at issue: confidence in the thinking done so far (which is the sort defined above) should reduce future search, while confidence in the benefit of additional thinking increases search (Baron, Badgio & Gaskins, 1986).

Excessively high confidence can be justified by further thinking, so long as its direction is fair. Thus, when someone is overconfident as a result of too little thinking, the two competing remedies are to reduce confidence or think more.

Importantly, the optimal amount of thinking to be done by individual citizens in matters of public policy may be low. For example, suppose you ask me how problems of the U.S. Post Office should be fixed. I can give you a few thoughts about it, and I could do a little reading and give you a few more. But, in the end, I will not be at all confident in any conclusion I could draw myself. In real

[4]It is normatively rational for evidence against a strong possibility to be weighed less favorably when the quality of the evidence itself is in doubt (Koehler, 1993; Baron & Jost, in press). For example, if an experiment supports precognition, we might justifiably suspect that something is wrong with the experiment. However, if AOT were formally represented, e.g., in a computer, this consideration would be applied to the description of the evidence itself rather than in the bias parameter.

[5]This statement assumes that the marginal benefit declines as search increases, while the marginal cost declines less or increases.

life, I would not think about this at all but would wait for some trusted expert to explain the issue to me (often in a publication).[6]

Confidence in a public-policy opinion should be low when little thinking, by anyone, has affected it. In many cases, extensive search is not worth the cost, and people rationally hold opinions that have not been subject to examination. But, according to the prescriptive view I have just sketched, they should know that they are doing this and thus refrain from confidently accepting their tentative conclusions as justified, and they should refrain from imposing them on others through their political action. They should follow the bumper-sticker maxim: "Don't believe everything you think." Many of the problems I sketched at the outset arise when overconfident politicians are trusted as authoritative experts by citizens who not only fail to apply AOT principles to their own thinking but also evaluate expertise in terms of expressed confidence — blanket, intuitively appealing, statements made without hesitation of qualification — rather than whether the assumed expertise is itself the result of AOT.

2.2 Biases that are common and general

When we look at how people actually think, we find a few systematic and general departures from this model of rational thinking. Many of these departures are biases that favor conclusions (possibilities) that are already favored. People tend to search selectively for evidence that favors these possibilities, whether they search for external information (Hart et al., 2009) or internal information based on memory (Perkins, 1985; Perkins, Bushey & Faraday, 1986; Baron, 1995; Gürçay-Morris, 2016). For example, Perkins et al. (1986) asked students to write down their thoughts on issues that were "genuinely vexed and timely" and that could be discussed on the basis of knowledge that most people have, e.g., "Would providing more money for public schools significantly improve the quality of teaching and learning?" Most students gave more arguments on their favored side, "myside" thoughts, than on the other side. When the students were asked to try harder to think of arguments on each side, they thought of very few additional myside arguments but many additional otherside arguments. Left to their own devices, then, many students looked primarily for reasons to support their initial opinion, but out of biased search rather than lack of ability or knowledge.

[6]The point of this example is different from the arguments for "rational ignorance" (Downs, 1957). The example would apply even if I were in a position of power, except that then I would invite experts to explain things rather than just waiting.

People also make inferences from evidence in a way that supports their pet conclusions (Lord, Ross & Lepper, 1979; Meszaros et al., 1996), even to the point of taking the same piece of evidence to favor different conclusions depending on which conclusion they favor (Baron, 2009). These biases together are called "confirmation bias" or "myside bias" in the literature.

In general, people search too little when search is warranted. We know this mainly from the fact that people who tend to search more do better in a variety of real-world manifestations of intelligence, such as school performance (Baron et al., 1986) and forecasting (Mellers et al., 2015). Yet the amount of search need not be correlated with fairness in the direction of search, nor with fairness of inferences, as I shall discuss.

Finally, confidence in judgments is generally too high when judgments are difficult. In most studies of the accuracy of confidence judgments, subjects are asked to provide answers to questions of fact plus a probability that their answer is correct. When the questions are difficult, the mean probability assigned by most subjects to a batch of questions is considerably higher than proportion that they answer correctly (Lichtenstein & Fischhoff, 1977). Controversial political issues such as immigration policy are often difficult in the relevant sense. AOT reduces overconfidence in difficult cases (Gürçay-Morris, 2016).

2.3 Individual differences in AOT

In an experiment closely related to those reported below, Baron (1995, Experiment 2; following up a similar study reported in Baron, 1991) asked subject to grade (from A+ to F) 24 lists of thoughts made about the morality of early abortion, supposedly written by other students in preparation for a class discussion. The subjects had previously written lists for themselves on the same topic. Each of the 24 list was composed of arguments for or against the proposition that early abortion is immoral. The number of arguments on the hypothetical student's side was either 2 or 4. The number on the other side, manipulated orthogonally, was 0 or 2. This was done for a hypothetical student on both sides, resulting in a 2x2x2 design. Each cell in this design had three lists, chosen to maximize the variety of the arguments used.

Most subjects gave higher grades to lists with arguments all on one side than to lists with arguments on both sides, even when the one-sides arguments opposed the subjects' own positions. Subjects who gave higher grades to one-sided lists were also the subjects who tended to give

arguments all on one side when making their own lists. Thus, judgments of the thinking of others, with respect to myside bias (direction of search), correlates with myside bias in one's own thinking.

Subjects also differed in the positive effect of number of arguments (amount of search) on their grades. Those who showed a larger effect of amount tended also to make more myside arguments in their own lists. But the effect of amount in judgment did not correlate at all with the number of other-side arguments they gave themselves. These results suggest that judgments of others' reasoning are affected by two somewhat uncorrelated dimensions, the amount of search and the direction of search (between fairness to both sides and myside bias).

This study shows that there is a relation between peoples standards for the evaluation of thinking, as applied to others, and their own dispositions. Can we measure these standards directly? Stanovich and West (1997, 1998) constructed a questionnaire that emphasized beliefs about the nature of good thinking, designed to measure beliefs that favored AOT. Importantly, they found that the AOT score correlated with several measures of actual task performance, i.e., manifestations of thinking dispositions. For example, they developed an Argument Evaluation Test to measure myside bias in the evaluation of arguments. Each of 23 items began with a fictitious person, e.g., Dale, stating an opinion about a social issue, for example, "The welfare system should be drastically cut back in size." The subject indicated agreement or disagreement (to indicate the subject's side). Dale then gave a justification, for example, "because welfare recipients take advantage of the system and buy expensive foods with their food stamps." A critic then presented a counterargument, for example, "Ninety-five percent of welfare recipients use their food stamps to obtain the bare essentials for their families." Finally, Dale rebuts the counterargument, for example, "Many people who are on welfare are lazy and don't want to work for a living." The subject then evaluated the strength of the rebuttal on a four point scale. The subject's answer was compared to answers given by experts — philosophy professors, and Stanovich and West. To estimate myside bias, the authors tried to predict each subject's ratings from both the expert ratings and the subject's own opinion about the issue. Myside bias was defined as a positive effect of the subject's beliefs. That is, subjects showing myside bias were those who tended to deviate from the expert ratings in the direction of their own prior opinions, rating arguments as better when they agreed with that opinion. Most subjects showed some myside bias, but some were more biased than others. The AOT score on the belief questionnaire correlated with myside bias on this test.

Other measures of beliefs have also correlated with various measures of the quality of thinking that ought to be affected by AOT as a disposition. Several papers (reviewed by Toplak, West & Stanovich, 2014, and Stanovich, 2016, Table 1), found correlations between various belief scales and other tests, some of which measured biases described in the literature on judgment and decision making, including (but not limited to): Baserate Neglect, Conjunction Fallacy, Framing Effects, Anchoring Effect, Sample Size Awareness, Regression to the Mean, Temporal Discounting, Gambler's Fallacy, Probability Matching, Overconfidence Effect, Outcome Bias, Ratio Bias, Ignoring P(D/ H), Sunk Cost Effect, Risk/Benefit Confounding, Omission Bias, Expected Value Maximization, Hindsight Bias, Certainty Effect, Willingness to pay/Willingness to accept, and Proportion Dominance Effect.

I selected items from the "flexible thinking" sub-scale of the original Stanovich/West questionnaire (the part most directly intended to measure AOT) and added a couple of others to make a short form appropriate for the general population, designed to assess beliefs in particular (first published, to my knowledge, by Haran, Ritov & Mellers, 2013).[7] Example items are: "People should take into consideration evidence that goes against their beliefs."; and "Changing your mind is a sign of weakness."[8] The short form of the scale has also had some success in predicting the results of other tasks that ought to be affected by AOT dispositions, such as perceptual judgments and reduced over-confidence in them (Haran et al., 2013), accuracy in geo-political forecasting (Mellers et al., 2015), utilitarian moral judgment, and problem solving (Baron, Scott, Fincher & Metz, 2015).

The fact that belief measures correlate with task performance suggests that efforts to explain to people the value of AOT, thus changing their beliefs to make them more favorable toward AOT, could result in improved performance on many tasks that involve thinking. Gürçay-Morris (2016) attempted to do this with a short training module, with some short-term success in reducing myside-bias and overconfidence on difficult problems.

More generally, one way in which people could come to have AOT dispositions is that they try (with varying success) to bring their own thinking into line with their standards for what good

[7] Svedholm-Häkkinen and Lindeman (2013) factor analyzed the long form, finding four factors, which they labeled Dogmatism, Fact Resistance, Liberalism and Belief Personification. My short form and the Stanovich/West flexible thinking subscale are similar to their Fact Resistance factor. The other three factors, and the other components of the Stanovich/West scale, represent traits that I would expect to correlate with AOT, but which are not essential to its definition. When such traits are included in the scale, we cannot ask empirical questions concerning the magnitude of their correlations with AOT beliefs as defined by central properties of AOT.

[8] A recent version of this scale is described later.

thinking is. Thus, we find people saying things like, "I may be biased in what I am about to say, but . . . ." Such statements indicate that they are at least recalling what their standards are. Surely such people are responsive when told, "Yes, you are biased. Here is why."

The acquisition of beliefs themselves surely comes in part from culture, including formal education. Particular cultures or subcultures maintain social norms consistent with AOT. That is, they reproach people who do not follow the norms, and they behave as if such people are doing something wrong (and possibly harmful to others). Thus, we have terms such as "pigheaded", "hasty", "ill-considered", "unthinking" and their opposites to express such norms to each other. Other subcultures may promote opposing norms, such as a respect for decisiveness, loyalty (to ideas), and respect for authority.

Suggestive support for the role of culture comes from Baron et al. (2015), who suggested that some people grow up in cultures that oppose questioning, lest children come to question doctrines dictated by authority. Baron et al. (2015) found large negative correlations between a measure of AOT and a measure of belief in "divine command theory" (Piazza & Landy, 2013), the idea that people do not have the capacity to engage in moral reasoning or to understand it, therefore, we must accept the word of God without question. Some cultural institutions, in order to prevent questioning of their authority and thus preserve the loyalty of their followers, may go so far as to inculcate the belief that thinking, curiosity, and questioning are more generally undesirable. To the extent to which this occurs, the promotion of AOT becomes part of a "culture war" rather than a technical problem.

In addition, people may come to understand on their own the elements of AOT as I have described them. In particular, they may come to understand that true knowledge is acquired only to the extent to which purported conclusions can survive attempts to knock them down. And they can see what is wrong with people who make overconfident, "cocksure" or self-assured statements about topic they know little about. And it can be clear to them how efforts to consider the other side can reduce such overconfidence. In particular, understanding of AOT could consist of the following beliefs:

- Correct conclusions are more likely when more than one possible conclusion is evaluated. Unless we consider other possibilities we cannot know if one is better than the front runner, the one we favor initially.

- Evaluations of possible conclusions will be more accurate when both positive and negative evidence is sought in a balanced way. If evidence is sought only for the front runner, we cannot find out if it is not the best conclusion.
- Evidence, once found and evaluated, should be used in a balanced way, that is, independently of whether it favors or opposes a the front runner. Otherwise we cannot discover that the front runner is not the best conclusion.
- Optimal decision making involves consideration of all relevant goals, not just the single goal that raised the issue initially. If we do not do this, we might end up subverting an important goal that would not be subverted by some other option.[9]
- Confidence in a favored conclusion should be high only when the thinking that reached it (possibly done by others) has involved all of the above: consideration of alternatives, balanced search for and use of evidence, and consideration of relevant goals for decisions. Without this, conclusions may be incorrect or not the the best ones available, and high confidence could lead to precipitous action and premature cessation of thinking.
- Confidence should thus be low with little thinking has been done or the thinking has been biased toward the front runner.

These are the beliefs that must be assessed with an adequate belief scale, but stated in a more colloquial way. Moreover, the issue is not so much whether people find these beliefs acceptable but whether they are sufficiently available so as to be applied to the evaluation of one's own thinking and that of others, and whether they are strong enough to overcome the force of contradictory beliefs derived from culture.

In sum, we have a plausible account of how *understanding* of AOT can manifest itself in beliefs about thinking, as measured by the AOT scale, beliefs that will in turn be applied to the thinking of others, and to ones own thinking. The latter application will push thinking dispositions in the direction of conforming to AOT as a standard. By this account, the belief scales can serve as a primary measure of AOT when we are looking for individual differences in those traits that AOT might affect.

[9]This principle is violated by current efforts of the Trump administration to ignore beneficial "side effects" in the cost-benefit analysis of environmental regulations, such as the (substantial) health benefits of regulation of greenhouse gas emissions from fossil fuels.

2.4 AOT as a set of dispositions

AOT beliefs, as measured by direct questions, may affect dispositions, and the judgments of others' thinking, as noted. Judgments of others' thinking are measured by grading tasks, in which subjects evaluate the thinking of others as revealed in various ways. It is less clear how to measure AOT dispositions in thinking itself. I suspect that a good test of "AOT dispositions" would not involve a single type of item, as AOT has implications for many different measurable properties of thinking, most of which are affected by other factors as well as by the particular content of the items used. These could include:

- Measures of overconfidence in difficult problems assess the AOT prescriptions concerning confidence.
- Thought listing tasks of the sort describe above assess myside bias.
- Tasks, like the Argument Evaluation Test, that examine the sensitivity of evaluations of others thinking to agreement with the subject's position, measure myside bias in a different way.
- Pre-decisional distortion (e.g., Chaxel, Russo & Kerimi, 2013; Russo, Carlson & Meloy, 2006), in which a tentative decision leads to selective biases that strengthen the favored option over the period of deliberation, measure biased search for evidence and biased use (interpretation) of evidence.
- Failure to consider alternative hypotheses when testing hypotheses (Baron, Beattie & Hershey, 1988) can measure search for alternative possibilities.
- The illusion of explanatory depth, in which people think they understand something until they try to explain it, those forcing them to try to think of arguments that they failed to look for previously, may be sensitive to overconfidence in the absence of adequate evidence; Fernbach et al. (2013) show this is related to political extremism.
- Belief overkill (Baron, 2009), the tendency to interpret evidence as supporting a favored conclusion, thus allowing people to maintain the belief that all arguments point in the same direction, assesses biased interpretation of evidence.

Note that the much longer list shown to correlate with AOT (Stanovich, 2016) includes various biases in judgments that could plausibly be understood as affected by AOT, but it seems (to me)

more difficult with these tasks to specify which particular prescriptions of AOT are involved. It thus seems likely (to me) that the correlation of these tasks with AOT is more the result of third factors that also affects AOT (such as education or culture), and less the result of a direct effect of AOT dispositions on the task.

2.5 AOT vs. reflection/impulsivity (R/I)

Attempts to measure the relation between political beliefs and thinking dispositions have generally relied on measures of "reflection" such as the Cognitive Reflection Test (Frederick, 2005), as well as various questionnaire measures (Jost, 2017). These measures ought to be related to AOT, but they may also assess a somewhat different trait, *reflection/impulsivity* (R/I). Kagan, Rosman, Day, Albert and Phillips (1964) defined "reflective" children are those who choose to be careful at the expense of speed on problems that are difficult but ultimately soluble. The "impulsives" are the children whose answers are fast and inaccurate. The test is useful for prediction. Reflectives tend to be older, to score higher on IQ tests (even when the tests are timed), and to be less prone to disruptive behavior in the classroom (Messer, 1976).

In tasks of the relevant sort (difficult but soluble), it is often possible to find a positive correlation between response time (RT) and accuracy, despite the fact that any measure of overall performance would generally favor both high accuracy and *low* RT. To measure R/I, a simple method is to compute z(accuracy)+z(log(RT)), that is, the standardized score of accuracy plus the standardized score of log(RT). The logarithm removes most of the skewness in the RT distribution, thus preventing excessively long RTs from overwhelming everything else. A number of different measures of R/I correlate with each other (see Baron et al., 1986, for a review).

In many studies, it turns out that log(RT) alone is consistent across tasks and is sometimes just as useful a predictor as accuracy. For example, log(RT) of items from the Raven's Progressive Matrices, as well as accuracy on these items, predicts forecasting accuracy (Mellers et al., 2015). Baron et al. (2015) also reported both cross-task consistency of RT and correlations with measures of utilitarian moral judgment, (dis)belief in divine-command theory (as noted above), as well as AOT. Baron et al. (1986) report a positive correlation between latency and a measure of IQ, as well as evidence of consistency across tasks. More recently, Pennycook, Cheyne, Koehler and Fugelsang (2013) show that religious skeptics make fewer errors *and* spend more time in a logical task involving misleading

syllogisms. Moreover, Baron et al. (2015) argued that the Cognitive Reflection Test (CRT, Frederick, 2005), widely used to measure "reflective cognitive style", is a useful predictor of performance in other tasks mainly because it is a measure of R/I.

Importantly, although R/I is conceptually related to AOT, and correlated with measures of AOT, it is not the same. R/I is largely a measure of amount of search, period. It is related to AOT because any sort of search for counter-evidence will result in increased search. But the R/I measure does not measure the direction of search, just its amount. In laboratory problem-solving tasks, where subjects are scored according to correct answers, more search leads to higher scores. In real life, the optimal amount of search is often very little or none, even when it comes to thinking about public policies or political candidates. When search occurs, the fairness of its direction, and overconfidence in unjustified conclusions, may be the more important dimensions of individual differences, at least for politics.

3 AOT and current U.S. politics

We have a fair amount of somewhat indirect evidence that AOT is related to current political attitudes in the U.S. We might expect similar results in Europe, where nativist political parties similar to the U.S. Republican Party (at the time of writing, 2018) have substantial support, and in Muslim countries such as Indonesia, Iran and Turkey, where conservative Islamist parties vie for power with more moderate approaches. But few relevant studies in these countries have been done.[10] I say that the evidence is indirect because some of it does not attempt measure AOT directly but rather relies on measures that might be expected to correlate with AOT. And other evidence does not measure partisan preferences directly but, similarly, uses political attitude measures that might be expected to correlate with that.[11]

Several studies use the CRT. A typical result is that of Deppe et al. (2015; including the re-analysis by Baron, 2015). In several different samples, reflective thought, as measured by the CRT, was negatively correlated with social conservatism (as determined by the subjects themselves), but

[10] An exception is Yilmaz & Saribay, 2017, who found that training people to think "analytically" reduced conservatism in a task that involved comparing specific positions of others, in Turkey.

[11] Partisan preference, however, may not be the aspect of political attitudes that are most sensitive to AOT. My guess is that AOT would be more predictive of attitudes about such issues as immigration, climate change, abortion, and family planning, and perhaps not predictive of other attitudes that divide political parties, such as those toward regulation of business.

essentially not correlated at all with economic conservatism. Presumably, "economic" conservatism consists of support for small government, low taxes, and less regulation of business. Other studies have asked mostly about "conservatism" in general, so we might expect that both types of attitudes are relevant to the responses in these. Jost (2017) reviews several results showing small negative correlations between CRT scores and conservatism in general.

A few studies have used various versions of the self-report AOT scale (some along with versions of the CRT):

- Svedholm and Lindeman (2013) found a negative correlation with paranormal beliefs.
- Swami, Voracek, et al. (2014, Study 1) found a modest but significant negative correlation between a version of the AOT scale and belief in conspiracy theories.
- Pennycook et al. (2014) found a correlation of −.49 between AOT and religious belief (defined by a scale that included belief in the supernatural), and a smaller correlation between religious belief in the CRT (−.26, in a separate study with the same population).
- Baron et al. (2015, described above) found a correlation of −.61 of belief in divine-command theory with AOT, and a smaller correlation (−.32) with an extended version of the CRT.
- Kahan and Corbin (2016) included both short form AOT scale and a few direct questions about partisan identification, in a U.S. sample chosen to be representative of the U.S. voting population. The AOT results, not reported in the paper, were reported (from the raw data) by Baron (2017a). In this sample, the CRT did not correlate at all with conservatism, but the AOT scale correlated −.27 with conservatism (which is −.41 disattenuated based on reliability of the two measures).
- Svedholm-Häkkinen and Lindeman (2017, Table 4) found a correlation of −.44 between a measure of superstitious beliefs and the Fact Resistance sub-scale of the original scale, this sub-scale corresponding closely the short form I have used; in the same study, the correlation between superstitious beliefs and CRT was −.22.[12]

Thus, AOT may be a better predictor of current U.S. politics, especially insofar as political attitudes are affected by religious commitments, than R/I (as measured by the CRT).

[12]I thank Annika Svedholm-Häkkinen for sending me this result.

Integrative complexity, a measure applied to the analysis of verbal products (as described, for example, by Suedfeld & Tetlock, 1977) is conceptually close to AOT. The scoring system is based on two measures: differentiation and integration. Differentiation is the acknowledgment of multiple views or perspectives. It is essentially equivalent to the idea of fairness in AOT. Integration involves some sort of synthesis of the resulting views. It is more difficult to score. And, in fact, differentiation alone does much of the work in accounting for correlations between integrative complexity and other measures. Jost (2017) reviews several studies showing negative correlations between integrative complexity and conservatism. Of the four most recent (2014 and 2015) two showed no correlation and two showed a very clear correlation.[13]

3.1 The benefits of AOT for citizens themselves

AOT helps citizens think more effectively, in a number of ways. We can infer this from all the other results concerning correlations between AOT and effective problem solving, resistance to common biases, and its negative correlations with questionable beliefs such as beliefs in the paranormal.

Issues that citizens face tend to be ones with arguments for competing views, if only views about how to overcome the forces of inertia. Openness to arguments on different sides can make citizens more likely to change their mind in the direction of good arguments. Change need not be complete to be beneficial. A little doubt can be a good thing.

Nor does change need not be immediate. When we have thought about something long enough to have reduced confidence, we are more open to additional arguments. Lower confidence rationally increases the utility of additional information and may thus lead to change over time.

More generally, when confidence is low, people are more likely to engage in further search, and more likely to change their conclusion (Thompson, Prowse Turner & Pennycook, 2011). Although confidence may be excessively high when an initial intuition is strong, people can probably learn not to put so much trust in these initial intuitions.

AOT also ought to permit better cooperation between political factions. Successful negotiation, in general, usually involves trade-offs on several attributes, such as working hours and salary in the

[13]Tetlock (1986) points out that integrative complexity is higher when the issue under discussion involves conflict between competing goals, as is often the case for American "liberals". Thus, these results may depend on the sampling of issues in these studies, which is probably, but not necessarily, somewhat representative of the issues most under discussion in a given period. The sampling of issues is particularly important in any attempt to characterize conservatism in general with the use of stimuli based on real policy debates (Baron & Jost, in press).

case of labor negotiations (Bazerman & Neale, 1992). Ideally, each party gives up on those attributes that is of greater concern to the other party. Such "log rolling" (or "integrative bargaining") is more likely when the parties are aware of the weaknesses in their own original positions. Direct evidence for the benefits of AOT for negotiation, however, is lacking and worth collecting.

Similarly, AOT ought to reduce the polarization and fanaticism that often ties up political systems in knots. It is extremely unlikely that any political party or pressure group is absolutely right on every issue. Those who realize this are surely more willing to compromise.

If we are all affected by the thinking of other citizens, through their political behavior, we have reason to want and encourage each other to think well. In this sense, AOT should also function as a social norm, if not a moral norm. It may do no good to "blame" our fellow citizens for poor thinking. But it may still help to understand that, if they fail to be actively open-minded themselves and fail to apply AOT as a way to evaluate sources, they are harming others in ways that should be discouraged in public discourse.

3.2 AOT as norms for evaluation of sources

AOT involves a set of thinking dispositions, but, in those who understand it, it also provides a set of norms (standards) for the evaluation of anyone's thinking, including the thinking of others (Baron, 1993). Indeed, John Stuart Mill was perhaps the clearest 19th century advocate of what I am calling AOT. In *On liberty* (1859, ch. 2), he writes (as part of a longer argument): "The whole strength and value, then, of human judgment, depending on the one property, that it can be set right when it is wrong, reliance can be placed on it only when the means of setting it right are kept constantly at hand. In the case of any person whose judgment is really deserving of confidence, how has it become so? Because he has kept his mind open to criticism of his opinions and conduct. Because it has been his practice to listen to all that could be said against him; to profit by as much of it as was just, and expound to himself, and upon occasion to others, the fallacy of what was fallacious."

Individual citizens do not have the time or background to delve deeply into policies concerning trade, immigration, crime or almost anything (Baron, 1993). Partly this is a function of the low expected-value of spending time informing ourselves, but even if we are passionately involved we cannot get to the bottom of all the issues we face. Too much is known for any one person to do this. We must rely on the conclusions of others, and we must be able to distinguish relatively trustworthy

sources from those that express gut-level intuitions as if they were proven facts or pearls of wisdom (Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2015, 2016).

For example, science, and many other forms of scholary inquiry (especially philosophy, these days), are based on actively open-minded thinking (AOT), refining themselves by challenging tentative beliefs. Astronomy differs from astrology because the latter has no standard procedures for thinking critically about its assertions. The same applies to a great deal of religious doctrine. Science, by contrast, engages in AOT at least as a group, if not within the heads of individual scientists. Scientists are rewarded (with publications, grants, promotions, jobs) for finding problems with the conclusions of other scientists. Individual scientists also try (perhaps not always hard enough) to anticipate possible criticisms before they try to publish something. These practices make science effective in approaching truth and understanding ever more closely.

As consumers of news, citizens would also do well to pay attention to signs of AOT. In particular, the more trustworthy news sources typically indicate the nature of their evidence and the extent to which it is trustworthy. They indicate (when possible) who the sources were, how many were consulted (even if anonymous), and whether or not a story was confirmed, and the extent to which the confirmations were independent. Sources of truly "fake news" usually do not provide such signals, and if they do they may be making it up (something that is harder to detect without consulting other sources).

Likewise, the application of the norms of scholarly inquiry, including AOT, in government itself can improve its effectiveness (Sunstein, 2017). It would be helpful if citizens understood the value of these advances.

AOT is not just about "critical" thinking insofar as that term suggests a skeptical attitude, possibly leading critical thinkers to doubt even when they should trust. The understanding of AOT leads to trust insofar as trust is warranted, and this need not involve looking for flaws, as long as we know that others have done so on our behalf.

Yet, if politicians make confident promises that they surely cannot keep and brag about their successes by self-assured assertions of dubious facts, they can sometimes gain power and keep it. This happens when citizens fail to apply the norms of AOT to these pronouncements, including norms concerning justified confidence. The air of confidence is a false signal of true expertise and is not recognized as such. Good citizenship requires the application of AOT norms, and citizens who

do not take these norms to heart are putting others in danger.

4 Measurement of AOT as judgments

Given the importance of the use of AOT for judging the thinking of others, it may be helpful to develop tests of exactly this function. Part of this understanding involves the determinants of appropriate confidence. Part of the problem in politics is that overconfidence in the absence of sufficient reflective thought is often taken at face value when it should be taken as a sign of poor thinking.

I report here two new but somewhat preliminary experiments[14] in which subjects judged other people's thinking (as in Baron, 1995). Both experiments include new versions of the short form of the AOT scale. The experiments themselves show that the scale predicts appropriate responses to expressions of confidence. And the methods used could suggest items for measurement of individual differences in such responses.

5 Experiment 1

In the first experiment, the subjects do not see the actual reasons that the hypothetical person provides but, instead, the subjects were just told how many of each the person thinks of. This strips away any possibility of subjects evaluating the reasons themselves. Again, the experiment manipulates pro- and con-reasons to examine their effect on evaluation.

5.1 Method

Subjects were 85[15] members of a panel who signed up for paid experiments (with payment set for each study, aiming at $12/hour). From past experiments, they were approximately representative of the U.S. in terms of income and education, but not sex: 31% were male. Ages ranged from 21 to 74 (median 47). Four other subjects were eliminated because they gave the same answer to all questions on every page.

[14] Available at http://finzi.psych.upenn.edu/~baron/ex/aot/t5.html and http://finzi.psych.upenn.edu/~baron/ex/aot/t8.html, with checks for complete responses turned off.

[15] I aimed for 100 but stopped the study after 24 hours, since no new responses were arriving.

The introduction read:

> This study is part of an attempt to develop a new measure of how people evaluate thinking. It also concerns judgments of confidence, expressed as a probability.
>
> You will go through 11 problems and give grades to someone else's thinking, which is revealed in several steps (all shown on the same page). Each of these pages has 10 questions total. The 11 problems differ in the type of reasoning they involve, and this is of primary interest.
>
> The thinking is described abstractly, in terms of reasons for or against some conclusion. Suppose that all reasons are equally good.
>
> Finally you will answer an 11-item scale concerning your views of what good thinking is. (This is a new version of a scale you may have answered before.)
>
> It is important to bear two things in mind: You are evaluating someone else's thinking, not her conclusions. And, because of this, your grades should not depend on whether you agree with her conclusion or not. In most cases, you won't know much about the conclusion anyway.

The items concerned 11 different topics: a doctor making a diagnosis, a moral decision by a manager, a scientist deciding how much his theory is supported, an appellate judge ruling on an appeal, a news reporter trying to determine who is responsible for a proposed policy, a consumer buying a TV, a voter deciding on a proposition supported by his party, an art collector considering a new painting, an investor deciding about a start-up, a woman deciding whether to accept a marriage proposal, and a college student questioning his family's religion. The order was randomized for each subject. Here is an example of the questions asked for one of the topics:

> An investor is trying to decide whether to invest in a start-up. The start-up looks good to him, so he is 80% sure that it will be a good investment.
>
> He tries to think of additional reasons in favor of this investment. He thinks of three of them. [G1][16] What grade would you give to this thinking, if thinking were to stop at this point?
> A+ A A- B+ B B- C+ C C- D+ D D- F
> [P1] What level of confidence (%) is appropriate if thinking were to stop at this point? (Choose the number closest to what you think.)
> 40 50 60 65 70 75 80 85 90 95 99 100
>
> He then tries to think of reasons against this investment. He thinks of one of these. [same two questions: G2 and P2]

[16]The subject did not see the letters G1–5 and P1–5, which are used here to refer to the questions.

The investor next looks for more reasons on both sides and finds one more reason on each side. What grade would you give to this thinking, if thinking were to stop at this point (after spending equal effort, overall, looking for arguments on both sides)? [G3, P3]

Now suppose that, from the start, the investor tries to think of reasons on both sides and succeeds in finding 4 reasons on the side initially favored and none on the other side. [G4, P4]

Finally suppose that, from the start, the investor tries to think only of reasons on the side initially favored, and finds 4 of them. [G5, P5]

The situations were thus:

1. 3 pro, 0 con, after search for pro only
2. 3 pro, 1 con, after additional search for con only
3. 4 pro, 2 con, after additional search for both sides
4. 4 pro, 0 con, after search for both sides
5. 4 pro, 0 con, after search for pro only

The grading scale was converted to 0 (F) to 12 (A+).

The Aot scale, answered on a 1–5 (1="completely agree" to 5="completely disagree") was as follows:

1. Allowing oneself to be convinced by a solid opposing argument is a sign of good character.
2. People should take into consideration evidence that goes against conclusions they favor.
3. Being undecided or unsure is the result of muddled thinking. (R)
4. People should revise their conclusions in response to relevant new information.
5. Changing your mind is a sign of weakness. (R)
6. People should search actively for reasons why they might be wrong.
7. It is OK to ignore evidence against your established beliefs. (R)
8. It is important to be loyal to your beliefs even when evidence is brought to bear against them. (R)

9. When we are faced with a new question, the first answer that occurs to us is usually best. (R)

10. [Good thinking leads to uncertainty when there are good arguments on both sides.]

11. When faced with a new question, we should consider more than one possible answer before reaching a conclusion.

(R) indicates reverse scoring. The mean score was .89 on the −2 to 2 scale (s.d., .60), and the reliability coefficient (α) was .75.[17]

5.2 Results

TABLE 1: Correlations and means of variables in Experiment 1. Grades (G1, G2, ...) are on a 0-13 scale. P's indicate confidence judgments as probabilities.

	Aot	G1	G2	G3	G4	G5	P1	P2	P3	P4	P5
G1	−0.10										
G2	0.06	0.67									
G3	0.18	0.48	0.84								
G4	0.24	0.34	0.51	0.55							
G5	−0.17	0.78	0.53	0.29	0.34						
P1	0.03	0.73	48.0	0.31	0.30	0.61					
P2	0.17	0.45	0.64	0.52	0.30	0.35	0.75				
P3	0.23	0.37	0.64	0.67	0.42	0.31	0.65	0.88			
P4	0.35	0.24	0.37	0.39	0.70	0.30	0.53	0.57	0.70		
P5	−0.03	0.57	0.32	0.13	0.29	0.66	0.79	0.57	0.43	0.35	
Means	0.89	7.77	8.05	8.85	9.37	7.34	0.76	0.76	0.80	0.83	0.77

$r > .18$ for $p < .05$, 2-tailed, uncorrected; $r > .25$ for $p < .01$.

Table 1 shows the correlations and means for the basic measures.

I computed an index representing how grades (G1–G5) should be affected by AOT: $G_{aot} = (G2 + G3 + G4)/3 - (G1 + G5)/2$. Here, G2, G3, and G4 should get higher grades because the thinker looked for reasons on the con side, compared to G1 and G5, where the search is biased to pro-reasons only.

For the confidence probabilities (P1–P5), the comparable index was $P_{aot} = (P2 - P1) + (P3 - (P1 + P2)/2) + (P4 - (P1 + P2 + P3)/3)$. Here, confidence in P2 should be higher than in P1

[17]Item 10 was an experimental addition to the scale, an attempt to assess an understanding of the appropriateness of lack of confidence. Unlike Item 3, which had the same purpose, the experiment failed. Item 10 did not correlate with the other items, and did not predict anything it was supposed to predict in other studies as well as this one. It is thus not included, but the results reported here were qualitatively the same if it was.

because we now know that the 3 Pro reasons outweigh the 1 Con reason, after a first attempt at each. P3 should be higher still (4 pro vs. 2 con after additional search on both sides), compared to both P1 and P2, for the same reason. P4 should be highest, because it yields a strong result (4 pro, 0 con) after unbiased search. (It is not clear where P5 should fall, since it found 4 pro reasons, and we do not know how many con reasons the thinker would have found if she had tried.)

For comparison, I also computed $G_{r/i}$ a R/I measure for grades, based solely on the number of reasons: $G_{r/i} = (G4 + G5)/2 - G1 + G3 - G2$. Here G4 and G5 (4 pro reasons) are compared to G1 (3 pro reasons), and G3 (6 reasons, mixed pro and con) is compared to G2 (4 reasons, mixed pro and con).

The main results are as follows:

Averaging across items, the correlation between G_{aot} and Aot was .31 ($p = .005$). Thus, the Aot scale predicted the judgments that subjects made about someone else's thinking.[18]

The correlation between P_{aot} and Aot was .33 ($p = .002$). Thus, the Aot scale predicted the appropriateness of subjects' own confidence judgments.

Note that all of these measures (P_{aot}, G_{aot}, and Aot) had positive means, so that these correlations could result from a tendency to give extreme responses on any scale. To check this possibility, I reduced all measures of individual items to three levels (positive, zero [middle of the scale], negative: hence 1,0,−1) and recomputed the correlations just reported. They were both essentially unchanged and still significant at the same level or better.

The correlation between $G_{r/i}$ and Aot was .19 ($p = .089$), but $G_{r/i}$ was correlated .46 with G_{aot}. When Aot was regressed on both G_{aot} and $G_{r/i}$, only G_{aot} was a strong predictor (standardized coefficient .28, $p = .021$) and $G_{r/i}$ played essentially no role (.06, $p = .628$). Thus, although Aot and r/i are naturally (and empirically) correlated, the present results are largely specific to Aot.

These results were general across the 11 issues. G_{aot} and P_{aot} had reliability coefficients α of .97 and .93, respectively. Surely much of this generality is the result of the issues being presented together as part of one experiment. However, there were some differences. The lowest ratings for G_{aot} were for the art purchase and marriage. For the art purchase, one comment suggested that this was not an important decision because an art dealer could sell the painting if he didn't like it. I

[18] An earlier study (t3) did not find this result. However, in the earlier study, the thinker's confidence was fixed at a high level. So most subjects seem to have interpreted the grading questions as pertaining to whether the high confidence was justified. Thus, they paid attention primarily to the number of pro reasons.

expected that marriage would get a low score because it is possible that some confirmation bias is beneficial and that many subjects know this (Murray, Holmes & Griffin, 1996).

6 Experiment 2

The second experiment was an attempt to find judgments of other people's confidence using more concrete examples. The task was to make a prediction based on features. In particular, on each trial, the subject saw two political views of a hypothetical U.S. voter, and one of their tasks was to assess the probability that the voter would vote Democrat in the 2018 national election. The issues were designed either to point in the same direction or different directions. The idea was that, in the latter case, high confidence one way or the other would not be justified. As it happened, some of the positions were more predictive than others, so my analysis took this into account. There was of course no right answer, but it is still possible to ask how confidence depends on whether the two positions pointed clearly in the same direction or whether they conflicted. Importantly, the subject's first task of the subject was to evaluate the thinking, on a scale of A+ to F, of someone who was 90% sure she was correct, and of someone who was 60% sure. Aot would imply that high confidence should be more justified when the two cues pointed in the same direction. After grading the thinking, the subject gave her own probability that the voter would vote Democrat.[19]

6.1 Method

The 74 subjects were drawn from the same panel as used in Experiment 1, with some overlap; they were 34% male, and ages ranged from 21 to 78 (median 47). Four other subjects were eliminated because they gave the same answer to all questions on every page.

The introduction began:

> Political party judgments (t8)
>
> Voters often have beliefs that do not match the positions of major political parties. The questions here concern prediction of whether a U.S. voter will vote Republican or

[19] A similar study that is not reported (t7) was a predecessor to this one (t8) and showed similar statistically significant correlations with subject's overconfidence in more difficult cases. But its main purpose was to ask about biases in diagnostic reasoning, i.e., selection of new evidence. It failed to demonstrate the biases of interest, apparently because subjects had great difficulty understanding the instructions, so, as a result, individual differences in bias could not be assessed.

Democrat in the next national election, on the basis of information about her political beliefs. (She will not vote for a third party.)

Then you are asked to grade someone else's thinking based on her confidence, when she has just the information that you have.

Each page took the following form:

This voter:

favors increased admission of asylum seekers;

opposes universal health insurance for all, without discrimination based on pre-existing conditions.

If someone with just this evidence said that the probability of voting Democrat was 90%, or if she said that it was 10%, this would make her 90% confident that her best guess would be correct. How would you grade her thinking if she were this confident?
A+ A A- B+ B B- C+ C C- D+ D D- F

What if she said that the probability was 60% or that it was 40% (so that she would be 60% confident in her best guess)?
A+ A A- B+ B B- C+ C C- D+ D D- F

What probability would you assign, for voting Democrat, based on just this evidence?

There were four issues: "a ban on abortion after 20 weeks of pregnancy", "increased admission of asylum seekers", "withdrawal from the NAFTA (North American Free Trade Agreement)", "universal health insurance for all, without discrimination based on pre-existing conditions". Each page listed two issues. (Democrats were assumed to oppose 1 and 3 and favor 2 and 4..) For each of the 6 possible pairs of issues, all combinations of "favors" and "opposes" were presented on different pages. The result was 24 pages. Order of the pages was randomized for each subject.

The Aot scale was the same as in Experiment 1 (with a different version of item 10, which is still excluded from analysis).

6.2 Results

The main dependent measure was Gdiff, the difference (on the 0–12 grading scale) between the first, high-confidence, grade and the second, low-confidence, grade. The difference, as defined, should be higher (more positive, or less negative) when high confidence is justified. The main hypothesis is

that this effect will be greater in subjects high in Aot, because they are more sensitive to the relation between confidence and evidence.

The 24 cases could be classified into three groups — clear Democrat, clear Republican, and conflicted — but there was substantial variation in the mean probabilities that subjects assigned to each case within groups. Thus, for prediction of optimal confidence, I used the absolute difference from 50% of the 24 mean probabilities of the 24 cases.[20] I call this Easiness; it ranged from 0.1 to 23.0 (out of a possible 50).

Of primary interest are the sensitivity of Gdiff to Easiness, and the sensitivity of the subject's own Confidence (absolute difference of stated probability from 50%) to Easiness. These should correlate with Aot. I regressed Gdiff, and Confidence, on Easiness for each subject. The slope of this regression correlated with Aot (across subjects) for both measures: $r = .35$ for Gdiff ($p = .003$)[21] and $r = .45$ for Confidence ($p = .000$).

A second question is whether the mean intercept of these regressions is negatively correlated with Aot. That is, would higher-Aot subjects have lower Confidence for the most difficult possible case, at the point where Easy is 0? These intercepts were in fact negatively correlated with Aot: $r = -.254$ ($p = .035$) for Gdiff[22] and $r = -.309$ ($p = .007$) for Confidence. Thus, the Aot scale correctly predicts lower confidence when low confidence is most warranted, both in subjects' own confidence and in their evaluations of the confidence of others given the same evidence.

It may also be of interest that this effect on confidence for difficult cases is not a side effect of a general reduction in confidence, or a reduction in the understanding that high confidence can be appropriate when evidence is strong. The intercept when Easy is 50 (the maximum) was positively correlated with Aot for both Gdiff ($r = .354$, $p = .003$) and Confidence ($r = .443$, $p = .000$).

[20] Analysis using just the classification of clear vs. conflicting gave the same general results, weaker but still statistically significant.

[21] For the analysis of Gdiff, I eliminated 4 subjects who gave the same pair of grades to all 24 cases, and one additional subject who wrote, in comments, "I am assuming the top answer is for the top question and the bottom answer is for the bottom question." This subject's grades appeared to apply separately to the two policies. Summary measures of this subject's grading responses were extreme outliers. When this subject is included, all results remained significant at $p < .05$ two-tailed except as noted.

[22] If the one outlier subject is included, this result was significant at $p = .043$ one-tailed, but not significant two-tailed.

7 Conclusion

7.1 Implications for research

The studies described here, and many of those reviewed, have two general deficiencies when it comes to understanding the relation between AOT and politics. One is in the nature of correlational studies. It is usually possible that some plausible third variable can account for observed correlations. For example, AOT could correlate with the tendency to understand and follow instructions in experiments. This problem can be partially remedied by training studies such as that describe by Gürçay-Morris (2016), although they too may pose the problem of knowing whether any effects are the result of experimenter demand, or some side effect of the training.

A related issue is that subjects will see the task of judging others' thinking as closely related to the self-report AOT measure. As a result, their correlations might be inflated by their juxtaposed measurement in a single session. One solution is to look at correlations of measures from different studies, done for different purposes (as done by Baron, 2017b).

Yet, it seems that we can measure individual differences in endorsement of AOT using self-report questionnaires, although these may need to be refined so that they focus more on the direction of search, and the relation between confidence and the amount and fairness of the thinking done so far. We can also measure individual differences in how people judge other people's thinking.

The second general problem is that the designs and materials are abstracted and remote from the real situation. It is more difficult to do more realistic experiments that are also well controlled, although some studies have made significant progress (Bronstein et al., 2018), using realistic news articles as stimuli. One further step is to examine how these individual differences related to judgments of real politicians and of others who claim to be trustworthy authorities on political topics. How do we know when a politician, or a pundit, is a fake, a fool, not worthy of our trust? One extension would involve realistic cases, possibly those resembling news reports of scientific experiments or of politically relevant events, where the cues to AOT include quotations from skeptical scientists or statements of how reports were confirmed.

Another approach is to look at correlations of AOT measures, including both self-report questionnaires and grading of others' thinking, with judgments of real statements of the sort used by (e.g.) by Pennycook et al., 2015, 2016). We could use excerpts from speeches, and apply some

sort of content analysis to them in order to measure the extent to which they indicate AOT (e.g., the "differentiation" component of integrative complexity). Or we could create artificial examples, with and without signs of integrative complexity.

However, in real life, politicians and pundits do not usually indicate appreciation of the "other side" when they assert, with high confidence, facts that are widely accepted and well supported. Perhaps a more generally useful indicator of the absence of AOT is the high-confidence assertion of outlandish "facts", not widely accepted or well supported, without much acknowledgment of the existence of another side (or with unsupported disparagement of the other side as liars or conspirators). One hypothesis is that those who accept such "alternative facts" will also give less credit to qualifying or hedging expressions concerning facts that they do not accept. It may have been their insensitivity to the absence of such qualifiers that led to their acceptance of whoppers to begin with.

7.2 Implications for political action

In my judgment (perhaps biased), we have enough circumstantial evidence to assume that the absence of AOT, especially as a standard for evaluation of sources, contributes to the political problems in today's world. It is thus worthwhile to ask how to improve people's thinking, and how to put into practice those steps that would seem possible to take and that may make a differences. Although scholars should not jump too early into advocacy of action, neither should we wait too long.

If AOT is part of a culture war, should its advocates be actively open-minded about the other side? Should we try to be "balanced" in our discussion of alternative ways of coming to have beliefs? There are many manifestations of this question, e.g., the controversy about whether reflective classroom discussion about the theory of evolution must give some time, or equal time, to creationism. Does equal time amount to "false balance"?[23]

My answer is that the most important thing is to teach people to *understand* the arguments about why, and when, AOT is superior to other forms of reasoning. They must understand it as a "design" in Perkins' (1986) sense. That is, they must know its purposes (coming up with the best answer, with appropriate confidence), its structure (testing, and revising or replacing, tentative conclusions, maintaining appropriate confidence, and so on) and the arguments about why this structure serves

[23]See Koehler (2016) on the dangers of false balance.

the purposes. Students can understand something without accepting it, so this is not indoctrination in its pure form. Of course, we do know what the outcome will be: if we teach understanding of some concept and test this understanding (as we must do, if want to teach it effectively), then in fact more students will accept it. But we are applying our incentives to understanding, and acceptance is a beneficial side effect.

If the culture warriors from the other side challenge us, then we must argue with them respectfully, but firmly. Is AOT special in this way? Does instruction in physics and astronomy affect how people think about the cosmos, in ways that might conflict with religious doctrine? And, of course, we must ask why we should accept someone's conclusions if all the arguments for them come from intuition or from historical longevity.

Citizens do not need to be very "smart" in the usual sense to know when they do not know something, and to figure out which authorities are trustworthy, by understanding in detail what those authorities have done, or not done, to reach their conclusions. They do need to understand how good thinking works. But this is not so hard if you think about it.

We have no better way. Alternatives such as "faith", "the heart", or acceptance of the word of authority have no built-in mechanism for self-correction. If their conclusions are wrong, we have no way to know, and, therefore, we also have no way to know when they are right.

References

Baron, J. (1985). *Rationality and intelligence*. New York: Cambridge University Press.

Baron, J. (1991). Beliefs about thinking. In J. F. Voss, D. N. Perkins & J. W. Segal (Eds.), *Informal reasoning and education*, pp. 169–186. Hillsdale, NJ: Erlbaum.

Baron, J. (1993). Why teach thinking? – An essay. *Applied Psychology: An International Review, 42*, 191–237.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning, 1*, 221–235.

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York: Cambridge University Press.

Baron, J. (2009). Belief overkill in political judgments. (Special issue on Psychological Approaches to Argumentation and Reasoning, edited by L. Rips). *Informal Logic, 29*, 368–378.

Baron, J. (2015). Supplement to Deppe et al. (2015). *Judgment and Decision Making, 10*(4).

Baron, J. (2017a). Comment on Kahan and Corbin: Can polarization increase with actively open-minded thinking? *Research and Politics, 4*(1). http://dx.doi.org/10.1177/2053168016688122.

Baron, J. (2017b). Protected values and other types of values. *Analyse & Kritik, 39*(1), 85–100. http://dx.doi.org/10.1515/auk-2017-0005.

Baron, J. (2018a). Individual mental abilities vs. the world's problems. *Journal of Intelligence, 6*(2), 23. http://dx.doi.org/10.3390/jintelligence6020023.

Baron, J. (2018b). Social norms for citizenship. *Social Research, 85*(1), 229–253.

Baron, J., Badgio, P., & Gaskins, I. W. (1986). Cognitive style and its improvement: A normative approach. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 3*, pp. 173–220. Hillsdale, NJ: Erlbaum.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. congruence, information, and certainty. *Organizational Behavior and Human Decision Processes, 42*, 88–110.

Baron, J., Gürçay, B., & Metz, S. E. (2017). Reflection, intuition, and actively open-minded thinking. In M. Toplak & J. Weller (Eds.), *Individual differences in judgment and decision making from a developmental context*, pp. 107–126.. New York: Routledge.

Baron, J., & Jost, J. T. (in press). False equivalence: Are liberals and conservatives in the U.S. equally "biased"? *Perspectives on Psychological Science*.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284.

Bazerman, M. H., & Neale, M. A. (1992). *Negotiating rationality*. New York: Free Press.

Bronstein, M., Pennycook, G., Bear, A., Rand, D., & Cannon, T.. (2018). Reduced analytic and actively open-minded thinking help to explain the link between belief in fake news and delusionality, dogmatism, and religious fundamentalism. Available at SSRN: https://ssrn.com/abstract=3172140.

Chaxel, A-S., Russo, J. E. & Kerimi, N. (2013). Preference-driven biases in decision makers' information search and evaluation. *Judgment and Decision Making, 8*, 561–576.

Dasgupta, P. S., Erlich, P. R., et al. (2013). Pervasive externalities at the population, consumption, and environment nexus. *Science, 340*(6130), 324–328.

Deppe, K. D., Gonzalez, F. J., Neiman, J. L., Jacobs, C., Pahlke, J., Smith, K. B., & Hibbing, J. R. (2015). Reflective liberals and intuitive conservatives: A look at the Cognitive Reflection Test and ideology. *Judgment and Decision Making, 10*, 314-331.

Downs, A. (1957). *An economic theory of democracy*. New York: Harper and Row.

Fernbach, P. M., Rogers, T., Fox, C. R. & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science, 24*(6), 939–946.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 24–42.

Godfray, H. C. J., Beddington, J. R., Crute, I. R. , Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M. & Toulmin, C. (2010). Food security: The challenge of feeding 9 billion people. *Science 327*, 812–818. DOI: 10.1126/science.1185383.

Gürçay-Morris, B. (2016). *The use of alternative reasons in probabilistic judgment*. Doctoral Dissertation, Department of Psychology, University of Pennsylvania. http://finzi.psych.upenn.edu/~baron/theses/GurcayMorrisDissertation.pdf.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making, 8*, 188–201.

Koehler, D. J. (2016). Can journalistic "false balance" distort public perception of consensus in expert opinion? *Journal of Experimental Psychology: Applied, 22*(1), 24–38.

Jervis, R. (1976). *Perception and misperception in international politics*. Princeton: Princeton University Press.

Jost, J. T. (2017). Ideological asymmetries and the essence of political psychology. *Political Psychology, 38*, 167–208.

Kagan, J., Rosman, B. L., Day, D., Albert, J., & Phillips, W. (1964). Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs, 78* (1, Whole No. 578).

Kahan, D. M., & Corbin, J. C. (2016). A note on the perverse effects of actively open-minded thinking on climate-change polarization. *Research and Politics, 3*(4).

Koehler, D. J. (2016). Can journalistic "false balance" distort public perception of consensus in expert opinion? *Journal of Experimental Psychology: Applied, 22*(1), 24–38.

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality.

Organizational Behavior and Human Decision Processes, 56, 28–55.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*, 2098–2109.

Mill, J. S. (1859). *On liberty*. London: J. W. Parker & Son.

Mellers B. A., Stone E., Atanasov P., Roghbaugh N., Metz S. E., Ungar L., & Tetlock P. E. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*, 1–14.

Messer, S. B. (1976). Reflection-impulsivity: A review. *Psychological Bulletin, 83*, 1026–1052.

Meszaros, J. R., Asch, D. A., Baron, J., Hershey, J. C., Kunreuther, H., & Schwartz-Buzaglo, J. (1996). Cognitive processes and the decisions of some parents to forego pertussis vaccination for their children. *Journal of Clinical Epidemiology, 49*, 697–703.

Murray, S. L., Holmes, J. G., & Griffin, D. W. (1996). The benefits of positive illusions: Idealization and the construction of satisfaction in close relationships. *Journal of Personality and Social Psychology, 70*, 79–98.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition, 42*(1), 1–10.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making, 10*, 549–563.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2016). It's still bullshit: Reply to Dalton (2016) *Judgment and Decision Making, 11*, 123–125.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2013). Belief bias during reasoning among religious believers and skeptics. *Psychonomic Bulletin and Review, 20*, 806–811.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology, 77*(5), 562–571.

Perkins, D. N. (1986). *Knowledge as design: Critical and creative thinking for teachers and learners*. Hillsdale, NJ: Erlbaum.

Perkins, D., Bushey, B., & Faraday, M. (1986). Learning to reason. Final report, Grant No. NIE–G–83–0028, Project No. 030717. Harvard Graduate School of Education. http://finzi.psych.upenn.edu/~baron/aot/perkins1986.pdf.

Piazza, J., & Landy, J. F. (2013). "Lean not on your own understanding": Belief that morality is founded on divine authority and non-utilitarian moral judgments. *Judgment and Decision Making, 8*, 639–661

Russo, J. E., Carlson, K. A., & Meloy, M. G. (2006). Choosing an inferior alternative. *Psychological Science, 17*, 899–904.

Singer, P. (1982). *The expanding circle: Ethics and sociobiology*. New York: Farrar, Strauss & Giroux.

Stanovidh, K. E. (2016). The comprehensive assessment of rational thinking. *Educational Psychologist, 51*(1), 23–34.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*(2), 342–357.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161–188.

Suedfeld, P., & Tetlock, P. E. (1977). Integrative complexity of communications in international crises. *Journal of Conflict Resolution, 21*, 169–184.

Sunstein, C. R. (2017). Deliberative democracy in the trenches. *Daedalus, 146*(3), 129–139.

Svedholm, A. M., & Lindeman, M. (2013). The separate roles of the reflective mind and involuntary inhibitory control in gatekeeping paranormal beliefs and the underlying intuitive confusions. *British Journal of Psychology, 104*(3), 303–319.

Svedholm-Häkkinen, A. M., & Lindeman, M. (2017). Actively open-minded thinking: Development of a shortened scale and disentangling attitudes toward knowledge and people. *Thinking and Reasoning, 24*(1), 21–40.

Swami, V., Voracek, M., Stieger, S., Tran, U. S., & Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. *Cognition, 133*(3), 572–585.

Tetlock, P. E. (1986). A value pluralism model of ideological reasoning. *Journal of Personality and Social Psychology, 50*, 819–827.

Thompson, V. A., Prowse Turner, J. A. & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107–140.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Rational thinking and cognitive sophistication:

Development, cognitive abilities, and thinking dispositions. *Developmental Psychology, 50*(4), 1037–1048.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. New York: Cambridge University Press.

Yilmaz, O., & Saribay, A. (2017). Analytic thought training promotes liberalism on contextualized (but not stable) political opinions. *Social Psychological and Personality Science, 1*(7).