Received: 28 August 2020 | Revised: 2 February 2021 | Accepted: 23 February 2021

DOI: 10.1111/jtsb.12274

ORIGINAL ARTICLE

WILEY

Masters of suspicion: A Bayesian decision model of motivated political reasoning

Francesco Rigoli

Department of Psychology, City, University of London, London, UK

Correspondence
Francesco Rigoli, Department of Psychology, City, University of London, Northampton Square, London, EC1V 0HB, UK.
Email: francesco.rigoli@city.ac.uk

Abstract

Motivated reasoning occurs when judgements subserve motives that go beyond accuracy seeking. Substantial evidence indicates that motivated political reasoning is ubiquitous. This is hard to reconcile with computational theories (following Marr's terminology, theories describing the fundamental principles underlying a cognitive process) like Bayesian inference, because these rely on accuracy maximization. Hence, motivated political reasoning is often interpreted as violating computational principles. Here we propose a different view by offering a computational account of motivated political reasoning which relies on the notion of Bayesian decision (instead of Bayesian inference). The key idea is that utility maximization, and not accuracy maximization, drives political thinking. This implies that agents will tend to endorse judgements that serve their instrumental goals even when evidence in support is poor (though agents will still believe their judgements are the most accurate). In this framework, motivated political reasoning is not interpreted as violating computational principles, although its nature is now conceived as pragmatic (i.e., serving instrumental goals) rather than epistemic (i.e., seeking understanding). The paper presents a mathematical description of the theory and shows how this can help interpreting important

phenomena in political psychology such contextual priming, stereotyping, and displaced aggression.

KEYWORDS
Bayesian, computational, displaced aggression, motivated reasoning, political reasoning, stereotype

1 | INTRODUCTION

In political psychology, a fundamental research question is why people endorse some political beliefs and reject others. A related question is why some individuals accept one interpretation of political events and other individuals accept an alternative interpretation. Answering these questions requires understanding how political reasoning works. In keeping with an influential perspective in cognitive science (Marr & Poggio, 1976), political reasoning (like any other psychological process) can be explored at three different levels of analysis: (i) a computational level, probing the underlying functional principles and asking what the logic or function of political reasoning is; (ii) an algorithmic level, investigating the fine-grained representations and dynamics that realise political reasoning; and (iii) an implementation level, focusing on its physical underpinnings (e.g., neuronal activity). Here we investigate political reasoning by focusing on the computational level of analysis. In psychology and neuroscience, the predominant computational perspective proposes that the brain realises Bayesian inference (Knill & Pouget, 2004; Oaksford & Chater, 2007). According to this view, prior beliefs (built from previous experience) are integrated with novel experience to obtain inference. This approach implicates a key role for a motivation to be accurate, in other words a drive to afford estimates which are as close as possible to reality. Applying Bayesian inference to political reasoning, the theory predicts that individuals will embrace interpretations considered to offer the most accurate description of society and politics (Bullock, 2009; Gerber & Green, 1999; Grynaviski, 2006). However, when assessing this prediction empirically, evidence suggests that an accuracy motivation, and hence Bayesian inference, is insufficient to explain political thinking. For example, individuals are more prone to endorse beliefs about society which support their own interest or the interest of their own group (e.g., socioeconomic group, ethnic group, gender group, etc.), even when these beliefs fit poorly with reality (Bartels, 2008; Bobo & Kluegel, 1993; Gilens, 1999; Kinder & Sanders, 1996). Also, stereotype and prejudice often appear to be driven by convenience rather than by an attempt to describe reality accurately (Kunda & Sinclair, 1999; Pettigrew & Marteens, 1995).

To address this puzzling empirical evidence, influential contemporary theories (Jost & Amodio, 2012; Jost et al., 2009; Kim et al., 2010; Lodge & Taber, 2013; Taber & Lodge, 2016) highlight the notion of motivated reasoning (Kunda, 1990), which, following early philosophical and psychological perspectives (for a review, see Jost & Banaji, 1994), views political beliefs as arising from motivations beyond simple accuracy seeking. An influential account (the John Q, Public model; Kim et al., 2010; Lodge & Taber, 2013; Taber & Lodge, 2016) proposes that the interaction of two modes of information processing, controlled and automatic, determines which political belief will be endorsed. This model attributes a stronger influence to automatic processes such as emotions and prior attitudes, and only a secondary influence to controlled processes based on rational considerations of available evidence. Therefore, in most

circumstances political beliefs would emerge as the product of automatic forces, with deliberation coming into play only afterwards in the form of post-hoc rationalization. This perspective highlights that individuals are blind to the automatic processes that determine their political convictions: at a conscious level, they would believe that their reasoning describes accurately available evidence. Another highly influential proposal explains support for a certain political view as dependent on three distinct motivations (Jost & Amodio, 2012; Jost et al., 2009). The first one (analogous to an accuracy motivation) is epistemic, corresponding to a drive for understanding how society works and to evaluate it. The second is an existential motivation, leading individuals to endorse political beliefs perceived as better for managing social threats. Finally, a relational motivation would drive people to hold political interpretations which allow them to foster socialization and bonding, and to pursue the interest of their own group.

Contemporary theories of motivated political reasoning (Jost & Amodio, 2012; Jost et al., 2009; Kim et al., 2010; Lodge & Taber, 2013; Taber & Lodge, 2016) offer highly valuable insight. However, these models are not computational (i.e., they do not explore the underlying logic or function of a psychological process (Marr & Poggio, 1976)). Hence, whether political reasoning can be understood within a computational perspective remains to be established (we have already seen that the notion of Bayesian inference seems to fail to do so). The goal of this paper is to elaborate on previous theories of motivated political reasoning and develop a computational model of political reasoning. We will see that this model still relies on Bayesian principles, though not on Bayesian inference but on the notion of Bayesian decision (Rigoli, 2020); hence the model is referred to as Bayesian Decision Model of Political Reasoning (BDMPR). In the next section, the basics of Bayesian modelling are overviewed for unfamiliar readers. Then, the BDMPR is presented and applied to explain a variety of manifestations of political thinking. The last session discusses the model with respect to a broader set of issues.

2 | BAYESIAN MODELLING

Before introducing the BDMPR, it is useful to briefly overview Bayesian modelling in cognitive science (Bishop, 2006). The typical scenario involves an unknown variable that needs to be guessed based on some information: imagine a person trying to estimate the price of a house who can ask three friends an opinion on this price. In the person's mind, this problem can be represented by four variables: the true price of the house (HP), the first friend's opinion on this price (F1), the second friend's opinion (F2), and the third friend's opinion (F3). This scenario can be described by adopting the formalism of Bayesian networks (Bishop, 2006), where each variable is represented by a circle (Figure 1) (note that these are all interval variables, conventionally represented by circles; categorical variables can also be implemented, conventionally represented by boxes). Arrows describe probabilistic dependences among variables. In this scenario, the house price is assumed to influence the friends' opinions (i.e., the arrow goes from the former to the latter), and not vice versa. This reflects the fact that the friends' opinions actually depend on the house price, and not vice versa. The friends' opinion variables, but not the house price variable, are shaded in grey. This convention indicates that the friends' opinions are directly known, whereas the true house price is not known (it is a latent variable), but it needs to be estimated indirectly. How does this estimation work? The idea is that the person adopts Bayesian statistics to integrate the friends' opinions plus her own prior guess about the house price (e.g., the person's own initial guess might be £300000). This integration results in a posterior estimate of the house price, corresponding to the final guess (formally, this is the

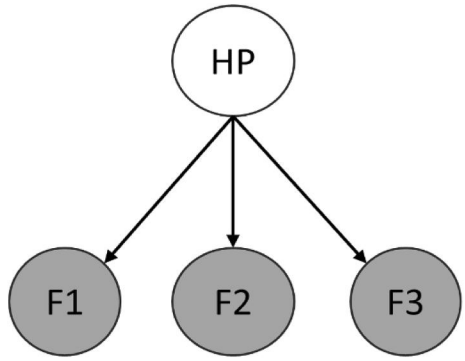

FIGURE 1 Bayesian network representing the scenario where a house price is estimated based on opinions from three friends. Its variables are: House Price (HP), first friend's opinion (F1), second friend's opinion (F2), third friend's opinion (F3). Arrows indicate probabilistic causal relations from one variable to another. Shaded variables are those considered to be observed

conditional probability of the house price given knowledge of the opinions of the three friends: P(HP|F1, F2, F3)). Importantly, in this framework, the person can attribute different reliability to different sources of information (formally, the reliability is captured by precision parameters; see Appendix). For example, the person might trust the first friend much more than the second. This entails that, when estimating the house price, the first friend's opinion will count much more than the second friend's opinion.

In short, Bayesian modelling offers a formal description of how latent variables can be estimated based on new information (the friends' opinion) and prior information (the own initial guess). These principles can be adopted to implement Bayesian inference, as in the example described here. When one of the variables captures a utility value, the same principles can be extended to implement Bayesian decision. Below, we explore how Bayesian decision can be applied to political reasoning, introducing the BDMPR.

3 | THE BAYESIAN DECISION MODEL OF POLITICAL REASONING (BDMPR)

The BDMPR is implemented by the Bayesian network represented in Figure 2 (a more formal description is offered in the Appendix). This describes the beliefs a social agent entertains about certain important variables of society and politics and about their relationships. The variables included in the model are represented by boxes (for categorical variables) and circles (for continuous variable). As above, arrows indicate probabilistic dependencies among variables. The first variable in the model is Hypothesis (Hyp), representing a categorical variable reflecting a set of mutually exclusive statements about society or politics. For example, one statement might claim that social benefits to unemployed people produce laziness (an anti-benefits hypothesis), and the alternative statement that social benefits foster job seeking (a pro-benefits hypothesis). These statements may not be evaluative, but simply descriptive (i.e., they may not imply any value judgement). Hyp plays a central role within the BDMPR, because the final result of the model is arbitrating among the different hypotheses implemented by Hyp. The second variable in the model is Prior Belief System (PBS). This represents a categorical variable reflecting a set of more general alternative views on society and politics. For example, one view

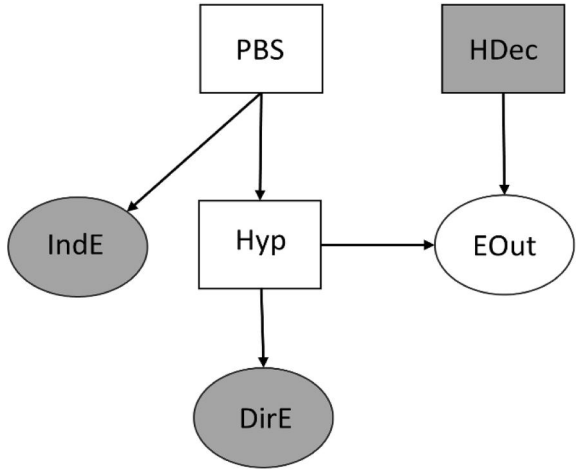

FIGURE 2 Bayesian network representing the model. Its variables are: Prior Belief Systems (PBS), Hypothesis (Hyp), Direct Evidence (DirE), Indirect Evidence (IndE), Hypothesis Decision (HDec), and Expected Outcome (EOut). Categorical and continuous variables are represented by rectangles and circles, respectively. Arrows indicate probabilistic causal relations from one variable to another. Shaded variables are those considered to be observed at each inference step

might be that state intervention promotes economic growth, and the alternative view that state intervention impairs growth (as for Hyp, no value judgement may be implicated). The variable Hyp depends on PBS, as the arrow going from the latter to the former indicates. For example, someone tending to view state intervention as promoting economic growth will also tend to attribute higher likelihood to the pro-benefits hypothesis.

In the model, both Hyp and PBS are treated as *hidden* (or *latent*) variables, as they cannot be observed directly but need to be inferred indirectly. For example, one does not know for sure whether state intervention promotes or impairs economy (PBS) nor whether social benefits encourage laziness or job seeking (Hyp). In addition to these two hidden variables (Hyp and PBS), the model includes variables that are observed, called Direct Evidence (DirE) and Indirect Evidence (IndE). The former is believed to be the consequence of Hyp and reflects novel information directly relevant for the specific hypotheses under consideration (e.g., a magazine article describing the impact of social benefits on employment). IndE is believed to be the consequence of PBS and reflects information relevant for the latter variable (e.g., a magazine article describing the impact of state intervention on economy), thus being also indirectly relevant for Hyp. Both DirE and IndE are represented by continuous variables (in our example, positive values for DirE correspond to evidence supporting the pro-benefits hypothesis, and positive values for IndE correspond to evidence supporting the hypothesis that state intervention promotes economic growth). Importantly, each evidence variable (DirE and IndE) is associated with a weight (formally, a precision parameter; see Appendix) which determines how influential that evidence is during inference. So far, we have considered PBS, DirE and IndE as single variables. However, a model might include multiple PBS variables. In addition, each PBS might project to multiple IndE variables and Hyp might project to multiple DirE variables. In these cases, each evidence variable could be associated with a specific weight (or precision) parameter, implying that each source of evidence will be more or less persuasive than the other. For example, one magazine might be considered as highly reliable, another magazine as extremely biased, a friend as trustworthy, another friend as deceitful, etc.

Finally, the BDMPR includes a Hypothesis Decision (HDec) variable and an Expected Outcome (EOut) variable. HDec is categorical and indicates which hypothesis of the variable Hyp is accepted as true and is used to guide behaviour. For example, HDec may include the following two categories: (i) accept the pro-benefits hypothesis (and support parties favouring social benefits) and (ii) accept the anti-benefits hypothesis (and support parties against social benefits). EOut reflects the expected outcome of this decision and depends on both Hyp and HDec. EOut is represented by a continuous variable where negative values correspond to punishment and positive values to reward. For example, EOut describes the outcome expected to occur (i) if the pro-benefits hypothesis is true and I accept it (and support parties favouring social benefits), (ii) if the anti-benefits hypothesis is true and I accept it (and support parties against social benefits), (iii) if the pro-benefits hypothesis is false but I accept it (and support parties favouring social benefits) (iv) if the anti-benefits hypothesis is false but I accept it (and support parties against social benefits).

The BDMPR realizes Bayesian decision by following a sequence of steps and eventually deciding which hypothesis to accept. Specifically, the model infers the consequences (in terms of reward or punishment EOut) of accepting different hypotheses considering evidence from DirE and IndE. Eventually, the hypothesis associated with the best consequence is accepted. More formally, this inference and decision process works as follows. DirE and IndE are observed and inference follows multiple steps. At each step, one different category of HDec is considered as observed and the posterior probability of EOut given DirE, IndE and HDec (i.e., P(EOut|DirE, IndE, HDec)) is calculated. This is repeated for all possible categories of HDec. After inference, decision follows, whereby the category of HDec associated with the best EOut (i.e., the highest posterior utility value) is chosen.

It is important to highlight that, in the BDMPR, the selected hypothesis is not necessarily the best supported by evidence (i.e., the one that maximizes accuracy), but the one associated with the best consequences (i.e., the one that maximizes utility). This emphasis on utility maximization distinguishes Bayesian decision theory from standard Bayesian inference. For example, the model predicts that an individual will be more likely to endorse the pro-benefits hypothesis if acceptance of this hypothesis is perceived as more advantageous compared to its rejection. Based on this reasoning, unemployed people are predicted to be more likely to endorse the pro-benefits hypothesis, because accepting, compared to rejecting, this hypothesis entails higher advantage for them. However, note that accuracy is still fundamental in the BDMPR. This is because accepting a hypothesis which is poorly supported by prior beliefs (PBS) and by evidence (DirE and IndE) is scarcely rewarding, implying that such hypothesis will be discarded.

According to the BDMPR, what is the phenomenological implication of accepting one hypothesis over the other? We propose that the implication is that, phenomenologically, an agent will believe that the accepted hypothesis is true even if, as explained above, it does not necessarily enjoy more support from evidence. In other words, the BDMPR postulates that agents are blind to the inference/decision process described above; they simply perceive the accepted hypothesis as true, without being aware that their perception is the product of utility maximization. This can explain the emergence of motivated political reasoning (Kunda, 1990).

In short, the BDMPR explains the genesis of political beliefs by relying on a Bayesian decision framework. The choice of such framework is motivated by the fact that such framework is computational and yet it can account for motivated reasoning (Kunda, 1990). It proposes that individuals consider prior belief systems together with novel evidence to infer the consequences of accepting alternative hypotheses, eventually endorsing the hypothesis

associated with the highest utility. This inference/decision process is postulated to be subconscious, and to ultimately result in the perception that the accepted hypothesis is true at the phenomenological level. This framework can help understanding the computations underlying some forms of motivated reasoning (Kunda, 1990). Crucially, in this view, motivated reasoning is not interpreted as being truly biased (i.e., something which violates computational principles), but only as apparently biased (given that ultimately it arises from the computational principle of utility maximization). Below, we will examine the role of each element of the model in the genesis of political thinking.

3.1 | The role of prior beliefs

By implementing the variable PBS, the BDMPR assumes that prior beliefs are critical in determining which political interpretation will be endorsed (Figure 3). Different forms of prior knowledge can be implemented in the model. First, prior beliefs can reflect knowledge about more general aspects. In the example above, while Hyp describes beliefs about a specific aspect of state intervention (i.e., regarding unemployment benefits), PBS captures beliefs about a more general influence of state intervention (i.e., regarding economic growth). Second, prior beliefs can reflect knowledge about a different, but partially related, context. For example, knowledge about the impact of maternity benefits can inform assessment of the impact of unemployment benefits. Third, prior beliefs can reflect accumulated experience at the personal level which is generalised to the whole society. For example, knowing how unemployment benefits impacted on the behaviour of a friend can inform assessment for the whole society.

We propose that prior beliefs in the BDMPR can explain contextual priming effects occurring during political thinking (Berger et al., 2008; Carter et al., 2011; Kalmoe & Gross, 2016; Todorov, 2005). In contextual priming experiments, participants are asked to report political judgements after being presented with apparently irrelevant information (a contextual prime). Strikingly, evidence has shown that this information, although apparently irrelevant, can bias political judgements (Berger et al., 2008; Carter et al., 2011; Kalmoe & Gross, 2016; Todorov et al., 2005). For example, after exposure to the American flag (a contextual prime), individuals are more likely to support the conservative party (Carter et al., 2011; Kalmoe & Gross, 2016). The BDMPR can account for these priming effects thanks to PBS and IndE. The latter captures evidence which, despite not being directly relevant for the hypotheses under consideration (Hyp), is directly relevant for PBS (Figure 4). According to the BDMPR, because Hyp and IndE both depend on PBS, information encoded by IndE will eventually exert an indirect influence over Hyp (Figure 4). As an example, consider a model where (i) PBS describes a general liberal versus conservative ideology, (ii) Hyp describes more specific beliefs about society (for example about the economy) which depend on the general ideology, (iii) IndE registers presence or absence of the American flag, and (iv) presence of the flag is treated as evidence supporting the conservative ideology. In line with empirical evidence (Carter et al., 2011; Kalmoe & Gross, 2016), when the American flag is shown the model predicts that beliefs about economy (implemented by Hyp) will be biased towards conservativism. This occurs because of the indirect influence of IndE upon Hyp via PBS.

In short, the BDMPR assumes that prior knowledge is critical in affecting which political belief will be endorsed. Also, the model highlights the role of evidence which is not directly relevant for the hypotheses under consideration, but which depends on prior beliefs. This

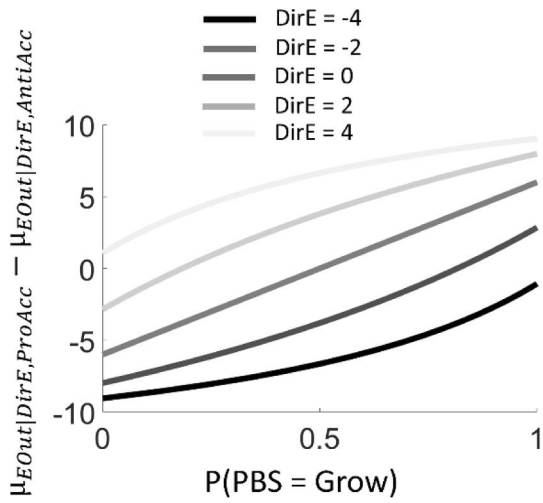

FIGURE 3 Description of the role of PBS. The simulated scenario is discussed also in the main text, where Hyp includes two categories (Pro-benefits hypothesis vs Anti-benefits hypothesis), PBS includes two categories (Grow vs NoGrow, reflecting whether state intervention promotes economic growth or not, respectively), positive values of DirE support the Pro-benefits hypothesis, and positive values of IndE supports the Grow category for PBS. The x axis reflects the prior probability for PBS = Grow. Different lines indicate different values for DirE. For all lines, IndE = 0, the precision parameter for DirE $\lambda_{SenE}{}^2 = 0.005$, the outcome of accepting the Pro-benefits hypothesis when it is true ($\mu_{EOut|Pro,ProAcc}$) is equal to 10, the outcome of accepting the Pro-benefits hypothesis when it is false ($\mu_{EOut|Anti,ProAcc}$) is equal to 0, the outcome of accepting the Anti-Benefits hypothesis when it is true ($\mu_{EOut|Anti,AntiAcc}$) is equal to 10, the outcome of accepting the Anti-benefits hypothesis when it is false ($\mu_{EOut|Anti,AntiAcc}$) is equal to 0. The y axis reflects the posterior outcome value of accepting the Pro-benefits hypothesis minus the posterior outcome value of accepting the Anti-benefits hypothesis

evidence is predicted to exert an indirect influence, a process that might underly some forms of contextual priming effects observed empirically (Berger et al., 2008; Carter et al., 2011; Kalmoe & Gross, 2016; Todorov, 2005).

3.2 | The role of direct evidence

Despite its emphasis on motivated reasoning, the BDMPR still views available evidence as highly relevant for political judgements. This notion is captured by the variable DirE (Figure 5), which describes evidence directly relevant for the hypotheses under considerations (Hyp). As an example, a magazine article discussing the impact of unemployment benefits will be considered as directly relevant for establishing whether benefits encourage laziness or job seeking. An important aspect is that the BDMPR associates a weight (or precision parameter) to the source of information (Figure 5). For example, if the article appears on a magazine perceived as poorly reliable (e.g., a magazine usually supporting a different ideology), it will be less influential on the final judgement. The notion of weight or precision becomes particularly relevant when a model includes multiple DirE variables, each associated with its own weight. This allows the BDMPR to capture the idea that different sources are imbued with different degrees of reliability. For example, a magazine generally supporting the own ideology will be attributed more weight (and thus will exert higher influence) compared to a magazine generally supporting a different ideology.

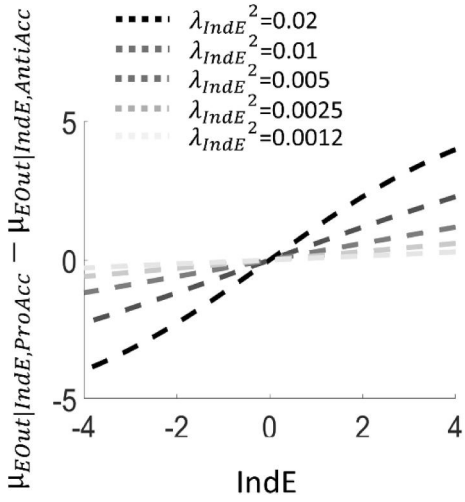

FIGURE 4 Description of the role of IndE. The simulated scenario is discussed also in the main text, where Hyp includes two categories (Pro-benefits hypothesis vs Anti-benefits hypothesis), PBS includes two categories (Grow vs NoGrow, reflecting whether state intervention promotes economic growth or not, respectively), positive values of DirE support the Pro-benefits hypothesis, and positive values of IndE supports the Grow category for PBS. The x axis reflects the value of IndE. Different lines indicate different values for precision parameter for IndE $\lambda_{IndE}{}^2$. For all lines, P(PBS = Grow) = 0.5, DirE = 0, the outcome of accepting the Pro-benefits hypothesis when it is true ($\mu_{EOut|Pro,ProAcc}$) is equal to 10, the outcome of accepting the Pro-benefits hypothesis when it is false ($\mu_{EOut|Anti,ProAcc}$) is equal to 0, the outcome of accepting the Anti-Benefits hypothesis when it is true ($\mu_{EOut|Anti,AntiAcc}$) is equal to 10, the outcome of accepting the Anti-benefits hypothesis when it is false ($\mu_{EOut|Anti,AntiAcc}$) is equal to 0. The y axis reflects the posterior outcome value of accepting the Pro-benefits hypothesis minus the posterior outcome value of accepting the Anti-benefits hypothesis

The role of available evidence DirE proposed by the BDMPR can explain empirical data on the influence of contextual ideological information (Cohen, 2003). One experiment presented participants with political statements usually associated with either right- or left-wing ideology (Cohen, 2003). Not surprisingly, left-wing participants tended to endorse left-wing statements, and right-wing participants right-wing statements. However, in one condition, while participants were presented with a statement, they were also informed about whether the conservative or the democratic party supported the statement. Strikingly, in this condition left-wing participants tended to endorse right-wing statements believed to be supported by the democratic party, and right-wing participants tended to endorse left-wing statements believed to be supported by the conservative party. In the context of the BDMPR, these findings can be understood by focusing on the role of DirE. The statement can be represented by Hyp (where categories indicate whether the statement is true or false), and whether the statement is supported by the conservative or democratic party can be represented by DirE. In other words, a participant could treat information about which party supports the statement as evidence relevant to establish whether the statement is true or not. For example, a left-wing participant would consider support of the democratic party as evidence that the statement is true. In this way, the BDMPR can explain empirical data showing that political judgements are highly dependent on contextual ideological information (Cohen, 2003).

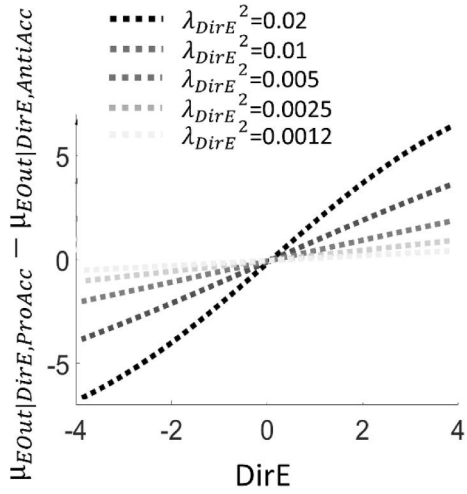

FIGURE 5 Description of the role of DirE. The simulated scenario is discussed also in the main text, where Hyp includes two categories (Pro-benefits hypothesis vs Anti-benefits hypothesis), PBS includes two categories (Grow vs NoGrow, reflecting whether state intervention promotes economic growth or not, respectively), positive values of DirE support the Pro-benefits hypothesis, and positive values of IndE supports the Grow category for PBS. The x axis reflects the value of DirE. Different lines indicate different values for precision parameter for DirE $\lambda_{DirE}{}^2$. For all lines, P(PBS = Grow) = 0.5, IndE = 0, the outcome of accepting the Pro-benefits hypothesis when it is true ($\mu_{EOut|Pro,ProAcc}$) is equal to 10, the outcome of accepting the Pro-benefits hypothesis when it is false ($\mu_{EOut|Anti,ProAcc}$) is equal to 0, the outcome of accepting the Anti-Benefits hypothesis when it is true ($\mu_{EOut|Anti,AntiAcc}$) is equal to 10, the outcome of accepting the Anti-benefits hypothesis when it is false ($\mu_{EOut|Anti,AntiAcc}$) is equal to 0. The y axis reflects the posterior outcome value of accepting the Pro-benefits hypothesis minus the posterior outcome value of accepting the Anti-benefits hypothesis

In short, despite an emphasis on motivated reasoning, in the BDMPR political judgements are still highly influenced by available evidence, as captured by DirE. The degree of influence exerted by evidence can vary for different sources, according to a weight or precision parameter. The role of direct evidence DirE in the BDMPR can potentially explain empirical findings indicating an influence of contextual ideological information (Cohen, 2003).

3.3 | The role of utility

The aspects examined so far (prior beliefs and evidence) are captured also by Bayesian inference (Knill & Pouget, 2004; Oaksford & Chater, 2007). The key aspect distinguishing the latter from Bayesian decision is the inclusion of a utility component. In the BDMPR, acceptance and rejection of any hypothesis is linked with an expected utility (Figure 6). Consider the example above comparing a pro-benefits versus anti-benefits hypothesis. Here the model asks: what is the consequence of accepting the hypothesis that benefits produce laziness (hence supporting anti-benefits parties) if the hypothesis is true? And if it is false? And what is the consequence of accepting the hypothesis that benefits promote job seeking (hence supporting pro-benefits parties) if the hypothesis is true? And if it is false? From the answers to these questions (and from considering how probable each hypothesis is), the model establishes whether, overall, accepting the pro-benefits hypothesis will be better (in terms of expected utility) than accepting

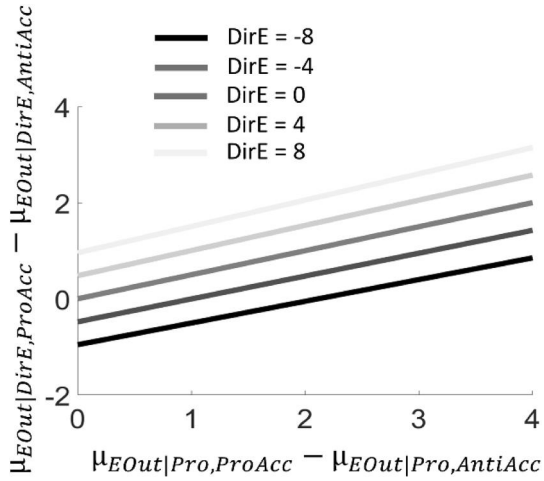

FIGURE 6 Description of the role of EOut. The simulated scenario is discussed also in the main text, where Hyp includes two categories (Pro-benefits hypothesis vs Anti-benefits hypothesis), PBS includes two categories (Grow vs NoGrow, reflecting whether state intervention promotes economic growth or not, respectively), positive values of DirE support the Pro-benefits hypothesis, and positive values of IndE supports the Grow category for PBS. The x axis reflects the difference between the expected outcome of accepting the Pro-benefits hypothesis when it is true ($\mu_{EOut|Pro,ProAcc}$) and the expected outcome of accepting the Pro-benefits hypothesis when it is false ($\mu_{EOut|Anti,ProAcc}$). Different lines indicate different values for DirE. For all lines, P (PBS = Grow) = 0.5, the precision parameter for DirE $\lambda_{DirE}{}^2 = 0.0012$, the outcome of accepting the Pro-benefits hypothesis when it is true ($\mu_{EOut|Pro,ProAcc}$) is equal to 20, the outcome of accepting the Pro-benefits hypothesis when it is false ($\mu_{EOut|Anti,ProAcc}$) is equal to 10, the outcome of accepting the Anti-benefits hypothesis when it is true ($\mu_{EOut|Anti,AntiAcc}$) is equal to 10. The y axis reflects the posterior outcome value of accepting the Pro-benefits hypothesis minus the posterior outcome value of accepting the Anti-benefits hypothesis

the anti-benefits hypothesis, or vice versa. The best hypothesis will then be accepted, and believed to be true at the phenomenological level.

An important aspect of this reasoning is that, according to the BDMPR, the accepted hypothesis will not necessarily be the one which is wished to be true, but rather the one considered to be more costly to reject. To understand this point, consider a conservative individual (i.e., someone valuing tradition as highly positive) arbitrating between the hypothesis that immigration is disrupting tradition against the hypothesis that immigration is not affecting tradition. For the individual, the hypothesis wished to be true will obviously be that immigration is not affecting tradition. However, which hypothesis is more costly to reject? The answer is that rejecting the hypothesis that immigration is disrupting tradition (implying no need to oppose immigration) will be more costly to reject. In this example, the model predicts that the individual will tend to accept the hypothesis that immigration is disrupting tradition (the one more costly to reject, not the one wished to be true). Notably, this prediction is specific to the BDMPR, whereas most previous models of political motivated reasoning and ideology claim that hypotheses wished to be true (and not hypotheses more costly to reject) are more attractive.

Another important aspect regards the definition of utility adopted by the BDMPR. Scholars often conceive utility as subjective, implying that individuals will evaluate different outcomes in their own peculiar way (Fishburn, 1970). This idea is also shared by the BDMPR. Many also interpret utility in terms of purely material and economic wealth (Fishburn, 1970). This approach is not followed by the BDMPR, where utility is viewed as encompassing multiple

forms of evaluation. With this regard, research has highlighted a set of core values which are qualitatively different, and which emerge in many different cultures (though different cultures will emphasise one or the other core value) (Schwartz, 1992). This perspective is relevant for the BDMPR, where utility may be conceived as an integration of multiple core values (with their relative weight varying across cultures and individuals). The reason for adopting this perspective on utility here is that a main argument ensuing from the BDMPR is that political reasoning is not purely "epistemic", but it is affected by one's own motivational values; what type of values one embraces is tangential to this argument.

Inclusion of a utility component allows the BDMPR to account for a set of empirical data. Substantial evidence shows that political beliefs depend on self-interest and on the interest of the own group (Bartels, 2008; Bobo & Kluegel, 1993; Gilens, 1999; Kinder & Sanders, 1996). The BDMPR proposes that this occurs because the hypothesis which promotes interests will be more costly to reject. Also, empirical evidence suggests that induction of death anxiety elicits ideological radicalization (i.e., liberals "become" more liberal and conservatives more conservative) (Castano et al., 2011; Kosloff et al., 2010). The BDMPR can explain this by proposing that experiencing psychological distress (such as death anxiety) will produce a change in the utility mapping. In other words, the proposal is that distress will magnify the distance in utility for rejecting/accepting different hypotheses. Consequently, the relative cost of rejecting the most appealing hypothesis will be even larger. In line with empirical evidence (Castano et al., 2011; Kosloff et al., 2010), this predicts that, in a state of death anxiety, the ideology previously accepted will be supported even more strongly. Note that this argument not only applies to death anxiety, but to any form of distress, something which remains to be investigated empirically.

It is worth noting that empirical literature has identified a more complex link between anxiety and ideology. Research has found a connection between higher trait anxiety and conservative ideology (Jost et al., 2003, 2007). In the context of the BDMPR, we suggest that this link is not dependent on the utility component, but on prior beliefs (implemented in PBS). Consider the belief (implemented in Hyp) that the economy is grounded on fierce competition, supported by the view (implemented in PBS) that humans are fundamentally egoistic and aggressive. These ideas are characteristic of many conservative ideologies. Viewing humans as egoistic and aggressive is likely to have implications not only for the political, but also for the personal, sphere. For example, it will increase the expectation of malignity in interpersonal relations. In turn, these expectations will be conducive of anxiety. In short, we suggest that the same prior belief that humans are egoistic and aggressive will support both conservativism in the political sphere and anxiety-inducing expectations in the personal sphere, hence explaining the empirical link between conservativism and high trait anxiety (Jost et al., 2003, 2007).

To summarise, our Bayesian decision proposal differs from Bayesian inference because of the inclusion of a utility component capturing forms of motivated political reasoning. This component can explain a tendency to endorse political beliefs which support self and group interests at the expense of accuracy. Also, the notion of utility offers an interpretation of why death anxiety increases radicalisation. The BDMPR predicts a similar effect is exerted by any form of distress, a prediction which remains to be tested empirically.

3.4 | Stereotype, displaced aggression, and anti-interest beliefs

Above, we have focused on specific elements of the BDMPR. Here we adopt a different approach and examine the model in the context of important phenomena in political

psychology. First, we focus on stereotyping (Bar-Tal et al., 2013). A stereotype is a belief about a social group which is retained despite contrary evidence. Stereotypes have been associated with prejudice (corresponding to a negative evaluation of the social group) and discrimination (Bar-Tal et al., 2013). To understand how the BDMPR explains stereotypes, consider the example of owners of land and of black slaves in 19th Century Southern USA. A common stereotype among these individuals was that black people were genetically different, for example more impulsive, less intelligent, and more violent. On this basis, slavery was viewed as ultimately beneficial for black people too, as it offered slaves a benevolent and paternalistic master. In the context of the BDMPR, we can treat this stereotype as corresponding to a hypothesis implemented by Hyp. The alternative hypothesis would be that black and white people are equal, implying that slavery is not beneficial for black people. Why did many landowners endorse the stereotype? The BDMPR proposes three factors at play. First, prior beliefs (PBS) might have been influential. For example, the belief that human races are genetically different in terms of psychological traits, and the belief that human society needs to be organized in hierarchies, were prior beliefs which supported the stereotype. Second, certain conditions were interpreted as evidence (DirE) supporting the stereotype hypothesis. For example, behaviour of poorly educated, maltreated and despised black slaves often resulted in impulsiveness, poor intelligence, and violence. These were interpreted as genetic predispositions, although in fact they were the very consequence of slavery (which paradoxically was proposed as the solution). A third aspect of the BDMPR which can explain stereotyping is the role of expected costs and benefits. In other words, at a subconscious level a landowner might have asked: what is the consequence of rejecting the stereotype hypothesis if it is true? And if it is false? And of rejecting the alternative hypothesis if it is true? And if it is false? Overall, rejecting the stereotype hypothesis might have appeared as much more costly, because it entailed criticising an advantageous social and economic system.

Displaced aggression (Dollard et al., 1939; Marcus-Newhall et al., 2000; Miller et al., 2003) in social and political contexts is another process that can be fruitfully interpreted by the BDMPR. Consider Germany right before Hitler took power. Because of the dramatic economic crisis exploded few years earlier, many people were experiencing the hardship of being jobless. We can imagine that people were seeking an explanation for their unemployment. Adopting the BDMPR, we speculate that two alternative explanations were available (implemented in Hyp). The first one is a systemic hypothesis, viewing the cyclic nature of capitalistic economy as responsible. The second hypothesis relies upon Jews' avarice in financial speculation. History tells us that many embraced the latter hypothesis, unleashing Hitler's unprecedented aggression towards Jews. Why did so many believe in Jewish guilt? Like with stereotyping, the BDMPR addresses this question by highlighting the role of prior beliefs. The idea of race as being genetically determined, and of races constantly fighting against each other, were prior beliefs which supported the anti-Jews hypothesis. Moreover, the BDMPR proposes a role for expected costs and benefits. At a subconscious level people might have asked: what is the consequence of accepting the anti-Jews hypothesis (and attacking Jews) if it is true? And if it is false? And of accepting the systemic hypothesis (and reforming capitalism) if it is true? And if it is false? Let us try to imagine how these questions were answered by many people. What is the consequence of accepting the anti-Jews hypothesis (and attacking Jews) if it is true? Many people might have answered that the consequence was highly positive. This is because Jews were a weak minority, and hence attacking Jews was predicted to be successful. Consider now the third question: what is the consequence of accepting the systemic hypothesis (and reforming capitalism) if it is true? Many people might have answered that the consequence was not very positive. This is because

reforming capitalism might have appeared as highly unlikely. Assuming an equal cost for wrongly accepting any of the two hypotheses, the BDMPR implicates that the anti-Jews hypothesis was more appealing. In other words, the BDMPR proposes that the anti-Jews hypothesis was more appealing because Jews were perceived as an easy target, while reforming capitalism was perceived as unlikely. Based on this cost/benefit reasoning, the anti-Jews hypothesis might have been eventually accepted despite being less supported by evidence (we note empirical data consistent with this interpretation; Berkowitz, 1959). This offers an example of how the BDMPR explains displaced aggression (Dollard et al., 1939; Marcus-Newhall et al., 2000; Miller et al., 2003) in social and political contexts, where initial suffering results in blaming a weak minority based on poor arguments, and in attacking the minority. Notably, the picture offered by the BDMRP is analogous to previous social psychology accounts such as the scapegoat theory of prejudice (Veltfort & Lee, 1943).

For a theory of political thinking based on motivated reasoning, it is critical to explain why individuals sometimes endorse beliefs which appear as going against their own and group interest. For example, why do many working class people vote conservative? Why do many people from disadvantaged ethnic minorities share stereotypes against their own ethnic group? Why do many females endorse sexist beliefs? When disadvantaged social groups (workers, ethnic minorities, females) endorse beliefs apparently going against their interest, we can talk about anti-interest beliefs. To examine how the BDMPR explains anti-interest beliefs, first it is important to emphasise that the beliefs described by the model are subjective. Therefore, some cases of anti-interest beliefs might simply be explained by failure to recognize certain hypotheses as less advantageous. For example, some working-class members might simply believe that cutting taxes for riches (a policy often supported by the conservative party) will not affect the welfare services they are relying upon. Or some religious peasants might have believed that revolting against a king blessed by God was going to lead to land reform but also to divine damnation.

However, there are circumstances where individuals are supposedly aware that a hypothesis is detrimental for them, and yet they accept it. For example, research has shown that many black people endorse racial stereotypes (Jost & Banaji, 1994; Katz & Braly, 1935; Sagar & Schofield, 1980) such as that black people are more violent. They are supposedly aware that this belief is detrimental for their interest, but they still endorse it. Why? The BDMPR proposes that this emerges because evidence (implemented by DirE) is interpreted as supporting the stereotype, counteracting any utility loss implicit in it. We can identify three aspects explaining why this occurs. A first critical aspect concerns how hypotheses are framed. For example, one individual might consider the hypothesis "black people are more violent" versus "black people are less violent". In this frame, evidence of black people's violence will be interpreted as supporting the first hypothesis. A second person might contemplate more nuanced hypotheses such as "violence depends on race" versus "violence depends on education and social status". Since education and social status are now considered, evidence of black people's violence will rarely support the first hypothesis (given that black people tend to be less educated and have lower social status; Jussim et al., 1987). Eventually, the first and second person will endorse and reject the stereotype, respectively. This raises the question of why hypotheses are framed the way they are in a society. This question goes beyond the scope of the paper, but socialization and power dynamics are likely to be critical. This implies that the dominant groups will have a primary role in shaping the categories used for political reasoning in a society (Gramsci, 1971; Jost & Banaji, 1994; Marx & Engels, 1846/1968).

A second aspect explaining why, in the context of the BDMPR, evidence is interpreted as supporting the stereotype (even when this goes against own and group interest) relies on the notion of self-fulfilling prophecy (Jussim et al., 1996). The latter consists in disregarding the fact that an event does not occur spontaneously, but it results from the very belief that the event occurs spontaneously. For example, the stereotype above might sometimes derive from not realising that black people's violence is in fact the very consequence of viewing black people as more violent. A third aspect is about the sources of information in a society. Our beliefs rely on evidence coming from two sources: from our own senses, and from other people or media. In large and complex societies, senses count less and the role of media is overwhelming. Media tend to be controlled by dominant groups, simply because these have larger economic and cultural resources. This implies that, on average, these media will be more likely to disseminate (even subtle) stereotypes against weaker groups than stereotypes against dominant groups (Mastro, 2009). In addition, media controlled by dominant groups will have more resources at their disposal, and hence better access to information. Therefore, their messages will be usually considered as more reliable (and will be weighted more during belief formation) also by disadvantaged groups. For these reasons, media will often expose a disadvantaged group to stereotypes that, despite being against the group's interest, are still treated as reliable by the group. In short, the BDMPR explains anti-interest beliefs by highlighting three factors (all linked with the role of evidence (DirE)): the way hypotheses are constructed, self-fulfilling prophecies, and the role of media.

To summarise, this section applies the BDMPR to a set of important phenomena in political psychology, including stereotype, displaced aggression, and anti-interest beliefs. This perspective offers new insight to further understand these phenomena. In addition, the model offers a unifying explanation of domains often viewed as separate, highlighting the common processes at play. Importantly, this is realised by reliance on the basic principles of Bayesian decision theory. We argue that a promising avenue for future research is to attempt applying this framework also to other phenomena in political psychology.

4 | DISCUSSION

This paper introduces the BDMPR, a theory of political thinking which extends the notion of Bayesian inference to the notion of Bayesian decision. While Bayesian inference models (Bullock, 2009; Gerber & Green, 1999; Grynaviskyi, 2006) struggle to explain motivated political reasoning, the latter is a critical element emerging from the BDMPR. By shifting from Bayesian inference to Bayesian decision, the theory views political reasoning as ultimately driven by utility, rather than accuracy, maximization. This implies that the nature of political reasoning is fundamentally pragmatic (i.e., serving instrumental goals) and not epistemic (i.e., seeking understanding). Critically, this implies that individuals are blind to the true motivations driving their beliefs. At the phenomenological level, the model assumes that individuals will sincerely believe that their conclusions are the best in light of evidence and prior knowledge. In other words, the model assumes some degree of self-deception. This assumption fits with the widely accepted view of motivated reasoning as acting at a subconscious level (Kunda, 1990). But is this also justifiable within a computational perspective (i.e., based on the general function and logic of a psychological process)? The following answer to this question can be proposed (Trivers, 2011). Within a computational perspective, political beliefs can be conceived as means to obtain social goals. To be effective, political beliefs would need to satisfy three fundamental requisites. First,

they would need to describe the social world accurately, an aspect the BDMPR captures by attributing importance to evidence and prior beliefs (if these are ignored, social goals will not be obtained). Second, they would need to take utility into account, also in line with the BDMPR. Third, because humans are primarily social animals, political beliefs will need to persuade others. Only if this occurs, political beliefs will ultimately be effective. In this perspective, self-deception might have evolved as an effective strategy to persuade others (a possibility which has received empirical support; Smith et al., 2017; Schwardmann & Van der Weele, 2019).

In the BDMPR, utility does not capture only material and economic conditions, but any form of value (the notion of core values could be useful to define utility in the context of the BDMRI; Schwartz, 1992). In other words, the BDMPR is agnostic about the nature of motivations underlying utility (i.e., it does not make any prediction about what motives drive political reasoning). With this regard, it has been suggested that epistemic, existential, and relational motives play a primary role in determining ideological convictions (Jost & Amodio, 2012; Jost et al., 2009). This view is compatible with the BDMPR, though the model fits also with the possibility that other motives (for example, purely material desires) are influential. Some scholars have proposed the existence of a specific motivation for justifying the current society (Jost & Banaji, 1994; Jost et al., 2019). Substantial evidence shows that individuals often believe that aspects of the current society are just, even if this goes against their interest (Jost & Banaji, 1994; Jost et al., 2019). However, whether this evidence reflects a specific motivation remains controversial (Jost et al., 2019; Mitchell & Tetlock, 2009). The BDMPR is compatible with the possibility of a specific motivation for justifying the current society. However, it also raises an alternative explanation: the belief that the current society is just might be a form of anti-interest beliefs (described above), which do not require any specific motivation for system justification. Note finally that the BDMPR can describe conditions where political reasoning is driven exclusively by accuracy seeking. These conditions occur when the cost of rejection/acceptance is equal across all hypotheses under considerations.

By adopting a computational approach (namely focusing on the function and logic of a psychological process) the BDMPR concerns a specific level of analysis. This approach is complemental to theories focusing on the algorithmic level, namely examining the fine-grained psychological processes at play (e.g., Kim et al., 2010; Lodge & Taber, 2013; Taber & Lodge, 2016). Some algorithmic models of political reasoning rely on the useful distinction between automatic and deliberative processes (Kim et al., 2010; Lodge & Taber, 2013; Taber & Lodge, 2016). The former are considered to be fast, rigid, and subconscious, the latter to be slow, flexible, and conscious. The BDMPR is suitable to describe automatic processes. However, the model can be extended to characterise also slower and more flexible processes acting at a subconscious level. This fits with the general idea that Bayesian accounts can be used to describe psychological processes at multiple levels, from perception to social cognition (Knill & Pouget, 2004; Oaksford & Chater, 2007).

The BDMPR operates at an abstract level (in Marr's terminology, at a computational level), and hence it offers a certain flexibility regarding how it can be fitted to specific problems. This flexibility characterises social science models operating at an abstract level, including Bayesian inference (Bullock, 2009; Gerber & Green, 1999; Grynaviskyi, 2006) and two-process models (Kim et al., 2010; Lodge & Taber, 2013; Taber & Lodge, 2016). Nevertheless, this does not imply that abstract models have no constrains and can explain every empirical phenomenon. For example, here we have argued that Bayesian inference models (Bullock, 2009; Gerber & Green, 1999; Grynaviskyi, 2006), despite their flexibility, are poorly equipped to explain motivated reasoning. Similar constrains apply to the BDMPR: for example, as described above, the

BDMPR proposes that hypotheses more costly to reject (and not hypotheses wished to be true) are more attractive. This prediction is specific to the BDMPR, whereas most previous models of political motivated reasoning and ideology claim that hypotheses wished to be true (and not hypotheses more costly to reject) are more attractive.

The paper focuses on psychological processes and not on the social dynamics that shape political reasoning. An important future research goal is to embed the BDMPR within a more general framework where political reasoning is shaped by social processes and in turn acts upon these social processes.

The BDMPR aims at offering a general theory of political reasoning, potentially valid in all contexts. General theories make necessary simplifications, and some instances might fit with the theory only poorly. However, their advantage is that they can offer insight on common principles underlying apparently different phenomena. The BDMPR relies on a mathematical modelling approach, which is relatively novel in the study of political reasoning. This approach implies some reductionism because it requires disregarding some subtle aspects of the concepts studied. However, it also offers a formal description of the processes involved, facilitating theoretical debate and formulation of empirical predictions. For this reason, mathematical modelling is becoming more and more popular in cognitive psychology and neuroscience; exploring its potentials in the context of political reasoning can potentially bridge research in cognitive psychology/neuroscience and research on social/political behaviour.

To summarise, the paper describes a novel theory which reconciles computational principles (specifically, the notion of Bayesian decision) with the concept of motivated political reasoning. We argue that the theory can inspire future research in at least two ways. First, the model can be used to interpret important phenomena in political psychology, and here we offer examples regarding stereotype, displaced aggression, and anti-interest beliefs. Second, by relying on a mathematical formulation, the model can be adopted to identify specific predications that can guide empirical investigation.

CONFLICT OF INTERESTS

The author(s) has/have no competing interests to declare.

ORCID

Francesco Rigoli https://orcid.org/0000-0003-2233-934X

REFERENCES

Bar-Tal, D., Graumann, C. F., Kruglanski, A. W., & Stroebe, W. (Eds.), (2013). *Stereotyping and prejudice: Changing conceptions*. Springer Science & Business Media.

Bartels, L. M. (2008). *Unequal democracy: The political economy of the new gilded age*. Princeton University Press.

Berger, J., Meredith, M., & Wheeler, S. C. (2008). Contextual priming: Where people vote affects how they vote. *Proceedings of the National Academy of Sciences*, *105*(26), 8846–8849.

Berkowitz, L. (1959). Anti-semitism and the displacement of aggression. *Journal of Abnormal and Social Psychology*, *59*(2), 182.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bobo, L., & Kluegel, J. R. (1993). Opposition to race-targeting: Self-interest, stratification ideology, or racial attitudes? *American Sociological Review*, *58*(4), 443–464.

Bullock, J. G. (2009). Partisan bias and the Bayesian ideal in the study of public opinion. *The Journal of Politics*, *71*(3), 1109–1124.

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science*, *22*(8), 1011–1018.

Castano, E., Leidner, B., Bonacossa, A., Nikkah, J., Perrulli, R., Spencer, B., & Humphrey, N. (2011). Ideology, fear of death, and death anxiety. *Political Psychology*, *32*(4), 601–621.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, *85*(5), 808.

Dollard, J., Miller, N. E., Doob, L. W., Mowrer, O. H., & Sears, R. R. (1939). *Frustration and aggression*. Yale University Press. https://doi.org/10.1037/10022-000

Fishburn, P. C. (1970). *Utility theory for decision making (No. RAC-R-105)*. Research Analysis Corp McLean VA.

Gerber, A., & Green, D. (1999). Misperceptions about perceptual bias. *Annual Review of Political Science*, *2*(1), 189–210.

Gilens, M. (1999). *Why Americans hate welfare: race, media, and the politics of antipoverty policy*. University of Chicago Press.

Gramsci, A. (1971). *Prison notebooks volume*. Columbia University Press.

Grynaviski, J. D. (2006). A Bayesian learning model with applications to party identification. *Journal of Theoretical Politics*, *18*(3), 323–346.

Jost, J. T., & Amodio, D. M. (2012). Political ideology as motivated social cognition: Behavioral and neuroscientific evidence. *Motivation and Emotion*, *36*(1), 55–64.

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, *33*(1), 1–27.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339.

Jost, J. T., Napier, J. L., Thorisdottir, H., Gosling, S. D., Palfai, T. P., & Ostafin, B. (2007). Are needs to manage uncertainty and threat associated with political conservatism or ideological extremity? *Personality and Social Psychology Bulletin*, *33*(7), 989–1007.

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual Review of Psychology*, *60*, 307–337.

Jost, J. T., Badaan, V., Goudarzi, S., Hoffarth, M., & Mogami, M. (2019). The future of system justification theory. *British Journal of Social Psychology*, *58*(2), 382–392.

Jussim, L., Coleman, L. M., & Lerch, L. (1987). The nature of stereotypes: A comparison and integration of three theories. *Journal of Personality and Social Psychology*, *52*(3), 536.

Jussim, L., Eccles, J., & Madon, S. (1996). Stereotypes, and Teacher Expectations: Accuracy and the Quest for the Powerful Self-Fulfilling Prophecy. *Advances in Experimental Social Psychology*, *28*, 281–388.

Kalmoe, N. P., & Gross, K. (2016). Cueing patriotism, prejudice, and partisanship in the age of Obama: Experimental tests of US flag imagery effects in presidential elections. *Political Psychology*, *37*(6), 883–899.

Katz, D., & Braly, K. W. (1935). Racial prejudice and racial stereotypes. *Journal of Abnormal and Social Psychology*, *30*(2), 175.

Kim, S. Y., Taber, C. S., & Lodge, M. (2010). A computational model of the citizen as motivated reasoner: Modeling the dynamics of the 2000 presidential election. *Political Behavior*, *32*(1), 1–28.

Kinder, D. R., Sanders, L. M., & Sanders, L. M. (1996). *Divided by color: Racial politics and democratic ideals*. University of Chicago Press.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Kosloff, S., Greenberg, J., & Solomon, S. (2010). The effects of mortality salience on political preferences: The roles of charisma and political orientation. *Journal of Experimental Social Psychology*, *46*(1), 139–145.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Kunda, Z., & Sinclair, L. (1999). Motivated reasoning with stereotypes: Activation, application, and inhibition. *Psychological Inquiry*, *10*(1), 12–22.

Lodge, M., & Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press.

Marcus-Newhall, A., Pedersen, W. C., Carlson, M., & Miller, N. (2000). Displaced aggression is alive and well: A meta-analytic review. *Journal of Personality and Social Psychology*, *78*(4), 670.

Marr, D., & Poggio, T. (1976). *From understanding computation to understanding neural circuitry*. Massachusetts Institute of Technology Press.

Marx, K., & Engels, F. (1846/1968). *The Germany ideology*. Progress Publishers.

Mastro, D. (2009). Racial/ethnic stereotyping and the media. In R. L. Nabi, & M. B. Oliver (Eds.), *Handbook of Media Processes and Effects* (pp. 377–391). Sage Press.

Miller, N., Pedersen, W. C., Earleywine, M., & Pollock, V. E. (2003). A theoretical model of triggered displaced aggression. *Personality and Social Psychology Review*, *7*(1), 75–97.

Mitchell, G., & Tetlock, P. E. (2009). Disentangling reasons and rationalizations: Exploring perceived fairness in hypothetical societies. *Social and Psychological Bases of Ideology and System Justification*, *1*, 126–158.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in Western Europe. *European Journal of Social Psychology*, *25*(1), 57–75.

Rigoli, F. (2020). A computational perspective on faith: Religious reasoning and Bayesian decision. *Religion, Brain & Behavior*, 1–18.

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in Black and White children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology*, *39*(4), 590.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in experimental social psychology*, *25*, 1–65. Academic Press.

Smith, M. K., Trivers, R., & von Hippel, W. (2017). Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, *63*, 93–101.

Schwardmann, P., & Van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, *3*(10), 1055–1061.

Taber, C. S., & Lodge, M. (2016). The illusion of choice in democratic politics: The unconscious impact of motivated political reasoning. *Political Psychology*, *37*, 61–85.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626.

Trivers, R. (2011). *The folly of fools New York*. Basic Books.

Veltfort, H. R., & Lee, G. E. (1943). The Cocoanut Grove fire: A study in scapegoating. *Journal of Abnormal and Social Psychology*, *38*(2S), 138.

How to cite this article: Rigoli F. Masters of suspicion: A Bayesian decision model of motivated political reasoning. *J Theory Soc Behav*. 2021;1–21. https://doi.org/10.1111/jtsb.12274

APPENDIX

Formally, the BDMPR is a mixture of Gaussians. The joint probability can be written as:

P(PBS, Hyp, HDec, EOut, DirE, IndE) = P(PBS) P(HDec) P(Hyp|PBS) P(IndE|PBS) P(DirE|Hyp) P (EOut|Hyp, HDec)

PBS is a categorical variable with number of categories equal to n_{PBS} and where each category is associated with a probability. If we consider the example about the influence of unemployment benefits (see above), we can set n_{PBS}=2, $PBS = Grow$ if state intervention promotes economic growth, and $PBS = NoGrow$ if state intervention impairs growth. The probability of promotion is $P(PBS = Grow) = x$ and the probability of impairment is $P(PBS = NoGrow) = 1 - x$ (where $0 \leq x \leq 1$). Hyp is also categorical, with number of categories equal to n_{Hyp}. Considering the same example, we can set $n_{Hyp} = 2$, $Hyp = Pro$ for the pro-benefit hypothesis (i.e., benefits encourage job seeking), and $Hyp = Anti$ for the anti-benefits hypothesis (i.e., if benefits encourage laziness). The conditional probabilities for Hyp are $P(Hyp = Pro \mid PBS = Grow) = y$, $P(Hyp = Anti \mid PBS = Grow) = 1 - y$, $P(Hyp = Pro \mid PBS = NoGrow) = z$,

$P(Hyp = Anti \mid PBS = NoGrow) = 1 - z$ (where $0 \leq y \leq 1$ and $0 \leq z \leq 1$). HDec is also categorical, with the number of categories $n_{HDec} = n_{Hyp}$. In our example, $HDec = ProAcc$ when the pro-benefits hypothesis is accepted (or, equivalently, when the anti-benefits hypothesise is rejected) and $HDec = AntiAcc$ when the anti-benefits hypothesis is accepted (or, equivalently, when the pro-benefits hypothesis is rejected). Probabilities for HDec are $P(HDec = ProAcc) = u$ and $P(HDec = AntiAcc) = 1 - u$ (where $0 \leq u \leq 1$).

DirE is a Gaussian variable conditioned on Hyp. Its conditional probability can be defined as:

$$P(DirE \mid Hyp = k) = \mathcal{N}\left(\mu_{DirE|k},\ 1/{\lambda_{DirE}}^2\right)$$

Here, every category of Hyp k has its own associated average $\mu_{DirE|k}$; for instance, the model will include $\mu_{DirE|Pro}$ (conditional on the pro-benefits hypothesis) which is different from $\mu_{DirE|Anti}$ (conditional on the anti-benefits hypothesis). The parameter ${\lambda_{DirE}}^2$ reflects the weight or precision of DirE and in the BDMPR it is equal for all levels of Hyp (in principle, a specific weight for each level of Hyp can be implemented). A similar logic applies to IndE with respect to PBS. The conditional probability is now:

$$P(IndE \mid PBS = i) = \mathcal{N}\left(\mu_{IndE|i},\ 1/{\lambda_{IndE}}^2\right)$$

Also for IndE every category of PBS i has its own associated average $\mu_{IndE|i}$; for instance, the model will include $\mu_{IndE|Grow}$ (conditional on the hypothesis that state intervention promotes economic growth) which is different from $\mu_{IndE|NoGrow}$ (conditional on the hypothesis that state intervention impairs economic growth). The parameter ${\lambda_{IndE}}^2$ reflects the weight or precision of IndE and in the BDMPR is equal for all levels of PBS (in principle, a specific weight for each level of PBS can be implemented).

Finally, EOut is a Gaussian variable conditioned on both Hyp and HDec. Its conditional probability is:

$$P(EOut \mid Hyp\ =\ k, HDec =\ j) = \mathcal{N}\left(\mu_{EOut|k,j},\ \sigma^2_{Eout}\right)$$

This indicates a specific average for each combination of Hyp and HDec. For instance, the model comprises $\mu_{EOut|Pro,ProAcc}$ (the expected outcome if the pro-benefits hypothesis is true and it is correctly accepted), $\mu_{EOut|Pro,AntiAcc}$ (the expected outcome if the pro-benefits hypothesis is true but it is wrongly rejected), $\mu_{EOut|Anti,AntiAcc}$ (the expected outcome if the anti-benefits hypothesis is true and it is correctly accepted), $\mu_{EOut|Anti,ProAcc}$ (the expected outcome if the anti-benefits hypothesis is true but it is wrongly rejected). The parameter σ^2_{Eout} reflects the uncertainty about the outcome and in our model it is equal for all combinations of Hyp and HDec (although in principle one can also implement a specific parameter for each combination).

The model is used to make inference. For inference, the variables DirE and IndE are observed, while the other variables are not. This includes multiple inference steps. At each step, for each level of HDec j, the model infers the conditional probability of EOut given the observed values for DirE and IndE and given HDec = j. This corresponds to the posterior Gaussian distribution:

$$P(EOut \mid DirE,\ IndE, HDec = j) = \mathcal{N}\left(\mu_{EOut|DirE,IndE,j},\ \sigma^2_{POST}\right)$$

Where $\mu_{EOut|DirE,IndE,j}$ is the posterior average for the expected outcome. For example, $\mu_{EOut|DirE,IndE,ProAcc}$ will be the posterior average if the pro-benefits hypothesis is accepted, and $\mu_{EOut|DirE,IndE,AntiAcc}$ is the posterior average if the anti-benefits hypothesis is accepted.

After all these inferences are made, the model makes a decision by choosing the hypothesis associated with the highest posterior $\mu_{EOut|DirE,IndE,j}$. For instance, it will either choose to accept the pro-benefits or the anti-benefits hypothesis.