AI & SOCIETY (2020) 35:177–185
https://doi.org/10.1007/s00146-019-00877-z

OPEN FORUM

The role of robotics and AI in technologically mediated human evolution: a constructive proposal

Jeffrey White[1,2]

Received: 30 January 2018 / Accepted: 26 December 2018 / Published online: 9 February 2019

Abstract
This paper proposes that existing computational modeling research programs may be combined into platforms for the information of public policy. The main idea is that computational models at select levels of organization may be integrated in natural terms describing biological cognition, thereby normalizing a platform for predictive simulations able to account for both human and environmental costs associated with different action plans and institutional arrangements over short and long time spans while minimizing computational requirements. Building from established research programs, the proposal aims to take advantage of current momentum in the direction of the integration of the cognitive with social and natural sciences, reduce start-up costs and increase speed of development. These are all important upshots given rising unease over the potential for AI and related technologies to shape the world going forward.

Keywords Cognitive social science · Computational model · Social simulation · Free energy principle · Directed evolution · AI arms race

1 Introduction

The potential for AI and related technologies to shape the future is increasingly an object of public and political concern. For instance, while addressing an audience of over 1 million on Knowledge Day, September 1, 2017, the President of the Russian Federation Vladimir Putin had this to say about AI and the future of humanity:

> Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world (as translated by Russia Today in RT 2017).

Whoever leads AI will rule the world. In the current political context, the tendency for many readers in English may be to interpret such a statement as a threat, or at least as a horsewhip in an arms race pitting state actor against state actor in a zero-sum effort to optimize artificial intelligence in the industries of war (cf. Armstrong et al. 2016).

However, there are other ways to see the future and the role of AI and robotics in shaping it (cf. White 2016). The purpose of this paper is to show that we need not anticipate an arms race eventuating in conflict, and rather that there is significant work ongoing in AI that points in the opposite direction. By recruiting and repurposing existing resources and established research programs, mutually beneficial ends may be identified, peaceful paths forward may be made explicit, and with these, rising anxieties due to currently resurgent geo-political polarization that might otherwise motivate a decision to initiate machine mediated conflict in resolution thereof may be quelled. Far from inviting mutual destruction, AI and related technologies may predicate open cooperation through thoroughly informed and mutually beneficial public policy, instead.

The next section introduces Sun et al.'s innovative integration of the social with the cognitive sciences, beginning with his proposal that cognitive social models may be essential for understanding cognition, generally, and then turning to his suggestion that the study of such models should be part of the curriculum of policy studies given their unique potential to make explicit to policy makers the indirect consequences of policy changes. Section 3 briefly introduces the

✉ Jeffrey White
jeffrey.white@oist.jp; j.b.white@utwente.nl

[1] University of Twente, Enschede, The Netherlands

[2] Okinawa Institute of Science and Technology, Onna, Japan

terms for translation between the cognitive, social and natural sciences required for the integration of models in a way that is informative to policy makers about policy impacts on individuals, communities and supporting ecologies, in the form of Friston et al.'s broad research program into the organizing principles of biological cognition. Section 4 briefly reviews three research programs in cognitive modeling, each at a selected level of organization complimenting the others in ways that, with results from one level informing the next, may provide for the testing of policy changes over short and long time spans. Section 5 briefly sketches how these three programs may be integrated in informing transitions through critical periods, and the paper ends by indicating research required for such predictive simulations to inform public policy.

2 The call to integration

In redress of prior schema offered by Newell and Simon (1976) and Marr (1982), Sun et al. (2005) proposed that integration across different levels of cognitive model may be necessary for an adequate understanding of cognition and attendant phenomena. Whereas predecessors focused on activity at different levels within individual cognitive agents, Sun et al. (2005) argued that any adequate account of intelligence must consider factors in terms of which intelligence emerges and in terms of which intelligent agents act, set out across four different levels of organization spanning (top to bottom): that accessible to sociological and anthropological inquiry including relationships between individuals and their environments, individual behavior as accessible to psychological inquiry, specialized modules and their assembly into functional whole brains accessible to cognitive science (as traditionally understood), and self-organizing physiochemical systems, i.e., living systems as accessible to fundamental biochemical inquiry (Sun et al. 2005, discussion pp. 619–621).

One reason given for Sun et al. (2005) development of this expanded schema is that, regardless of field, it is impractical for a single science to yield complete information at every level of organization at once. Rather, scientists working at one level of organization routinely rely on those at others to account for phenomena in terms outside of focal areas; and, in their discourse, understanding at one level is refined as it is checked against results from other levels, thereby improving the accuracy and predictive power of all. Sun et al. (2005) contend that we should expect the same dynamics to play out in the cognitive sciences, with models ultimately enriched to represent dynamics at increasingly higher and lower levels of organization in increasingly realistic terms. To increase the psychological realism of social simulations for example, Sun (2012) explicitly calls for the grounding of the social in the cognitive sciences. Most recently, Sun (2018b) calls for the "blending" (p. 245) of cognitive with social models as well, in order that their "integration" results in "tools for more precisely understanding policy implications at both individual and social levels" (p. 240) at the same time.

Specifically on the issue of the purposeful development of integrative models for the information of public policy, Sun (2018b) argues that computational models may benefit policy makers who instead of "relying on speculations" need a more "reliable means for understanding" policy implications (p. 240). Psychologically realistic models of social systems "may be used to predict human performance in organizational settings and to prescribe optimal or near-optimal cognitive abilities for individuals for specific tasks and organizational structures" (p. 243). Their development "for improved policy making" may take advantage of the "prior validation" of established work from which they are assembled, as demonstrated successes of component models "may be leveraged in validating the overall simulation results" (p. 244). Importantly, this transfer of validity should also extend to policy decisions made on the basis of given results, with the first step being the development of simulations tailored from established and ongoing research to this end.

In short, the selective integration of computational models may afford a privileged window on policy implications, and leveraging established programs do so in a reliable and timely manner. The practical issue for the cognitive social scientist becomes, then, understanding how programs at different levels of inquiry might fit together and in what terms their integration may take place so as to best inform this process. This is the subject of the next section.

3 Possible terms of integration

To weigh options, we need to account for both human as well as ecological costs due to a given policy or change and compare how they differ over time. To "blend" cognitive with social models to provide a means for this comparison, so that we may more precisely understand the implications of certain practices at individual, social and environmental levels all at once, terms of integration must be identified that facilitate their account in common.

One possibility exists in Karl Friston et al.'s recent work in Markov blankets (cf. Friston 2013; Cockshott and Renaud 2016; Ramstead et al. 2018) providing a framework for the translation between levels of description of human cognition in terms of Friston's "free energy principle" (FEP) (cf. Friston 2010, 2012). What is especially promising about this approach is that it accounts for cognition in terms of and as constrained by natural energetics beginning with Friston's

FEP, which for our purposes provides a conceptual bridge between changes in human costs due to policies supporting given institutional arrangements over relatively short time spans (typically intended to reduce costs to some human beings at the expense of their environments, including very often other human beings) and demands on supporting natural environments due those same policies and institutional arrangements over longer time spans.

Friston's free energy principle (FEP) formalizes cognition in terms of the maximization of expected utility, reward or value, through the minimization of prediction error, surprise, or cost, by way of "active inference" which involves maximizing evidence for internal models of the world as informed through ongoing sensory input. Generally, Friston et al. understand that neural structures within an organism work to minimize differences between anticipated ends and perceived results, with future intentions to act modified accordingly, and with the aim of this process being to secure the organism's present and future integrity against disintegrative change. The FEP expresses the key relationship between the cognitive agent's perceived and anticipated possible ends in terms of uncertainty, with the agent essentially motivated to avoid surprise. When anticipated ends match perceived results, the surprise is zero. When they do not, surprise demands attention and resources are expended. This is important given a scarcity of resources, and motivates agent psychology generally. Complicated internal models, political philosophies and economic systems are all expressions of cognitive systems operating according to this organizational principle, developing according to the implicit aim of minimizing uncertainty through the proactive organization of self, other and environment at the expense of energies collected and distributed through increasingly complex social arrangements.

Due to the explanatory scope of this program, it presents itself as providing possible terms for the integration of cognitive models required should policy informing simulations of the sort proposed by White (2016) and by Sun (2018b) be realized. Noteworthy in the present context is that Friston's FEP has already been employed in inquiries into cognition at different levels of organization.

- Ramstead et al. (2018) use the FEP to characterize cognition in terms of a general theory of dynamic systems consistent with evolutionary systems theory (EST). On this account, cognition is an aspect of living systems which maintain themselves in a limited number of stable states far from thermodynamic equilibrium with their environments by organizing themselves and their environments in such a way as to minimize disorder and surprise at the failure to deliver anticipated results, in effect securing preferred modes of "coupling" with their environments. "Consistent with EST, this propensity to minimize surprise is the result of natural selection …self-organizing systems that are able to avoid entropic, internal phase-transitions have been selected over those that could not" (Ramstead et al. 2018, p. 3).
- At the level of social organization in the context of economics, the free energy principle has proven more successful in understanding choice behavior as the minimization of surprise coupled with utility maximization than have other approaches which try to model the same phenomena in terms of utility maximization alone (Schwartenbeck et al. 2015).
- At the level of situated cognition, the FEP has recently been deployed in understanding how stress contributes to disease in organisms. Peters et al. (2017) interpret stress according to the free energy principle as "uncertainty" or "surprise" that frustrates the fundamental motivation to minimize entropy. In response, the brain taxes the body system by demanding more energy to rectify the condition and thereby diverting attention away from immediate tasks, with this "allostatic load" resulting in impaired memory, increased markers for cardiovascular disease, and diabetes.
- And in neuropsychology for example, Friston's FEP has been employed in the study of mirror neuron activity in the motor cortex in order to understand how brains switch between perception and action (cf. Shipp et al. 2013), as well as in accounting for reward learning as driven by dopamine excitation or depletion (Fitzgerald et al. 2015; see also Friston et al. 2016).

The next section introduces three different research programs that fall roughly into the evolutionary, social and neurodynamic levels of organization, before briefly sketching how they may all work together to advise public policy in Sect. 5.

4 Levels of model

This section reviews three active research programs in cognitive modeling, each at a different level of organization. The first is Peirera et al.'s evolutionary psychological approach employing logic programming to investigate the effects of different expressions of moral agency (apology, forgiveness, preconditions to cooperation, and so on) on group performance over evolutionary time. The second is Sun et al.'s cognitive social science approach focusing on psychologically realistic computational models of social intelligence. The third is Tani et al.'s neurodynamic approach grounded on predictive coding and directly demonstrative of Friston's FEP in learning neurorobots. The fifth section then sketches in general terms how these three may be integrated to assess human and natural resource costs of competing policy proposals.

4.1 Pereira's evolutionary game theory

L. M. Pereira et al. use logic programming and evolutionary game theory to model the capacity for individual agents to make moral decisions through abduction, either in reaction to contextual cues or through purposeful deliberation over points of interest, and from this basis have worked on understanding the roles of intention recognition, commitment, apology, forgiveness, revenge, ostracism, and guilt in cooperative collectives of similarly endowed individuals. The aim of this research is to better understand the emergence of cooperation as supported by cognitive mechanisms which thereby stabilize social orders over evolutionary time scales, so that this understanding may both inform human practice today, and so that such capacities may be interred in future robotic agents free to act within future human communities tomorrow. Thus, one strong theme running through Pereira et al.'s work has been the need to bridge individual with collective "realms" toward the goal of understanding just how individual agents act from the basis of one in the furtherance of the other (cf. Pereira and Saptawijaya 2015, 2016; Saptawijaya and Pereira 2018; also Han and Pereira 2018).

Pereira et al. account for native agent motivation to best available ends in terms of abduction. Historically, abduction has been variably given, depending on researcher and context. For Peirce—the inventor of the concept—abduction was variably characterized as well, depending on at which stage of his life one were to have asked him about it. According to a mature view, abduction is a natural tendency to discovery of truth, a "guessing instinct" through which (typically more successful than not) hypotheses are created and provisionally adopted as they are then tested through induction and clarified through deduction before contributing to further creative abduction (cf. Paavola 2006, discussion Chap. 4; also Aliseda 2006; Gabbay and Woods 2005; and Magnani 2017; see also Simon 1977, for an early view on abduction encoded as a computer program). In Pereira et al.'s models, abduction is a matter of determining a set of actions that satisfy goal conditions while maintaining personal integrity. These are iterated as "abducibles" and are evaluated by agents using doctrines of double and triple effect, utility functions, and counterfactuals for example (cf. Han and Pereira 2013; Han et al. 2015; Pereira and Saptawijaya 2017). The ethical norms of a group evolve as agents organize around increasingly ideal solutions afforded by increased agent-level capacities to cooperate, pursuing strategies for higher group-level payoffs of which agents share.

On Pereira et al.'s model, an agent's prior deliberate decisions to act in given situations are retained, and these can be employed reactively in similar future situations without the need to compute them again (cf. Saptawijaya and Pereira 2013; also Saptawijaya and Pereira 2018). Moreover, this "tabling" technique opens prior decisions to comparison between agents and allows for different agents to inform each other about differently determined optimal actions in different ways (cf. Pereira et al. 2013). Agents are also able to recognize intentions, to assess relative commitments to goals, to cooperate with each other where projected payoffs are better, and learn to coordinate actions only with those also prone to cooperation. With other prosocial capacities, such as the abilities to issue and to accept apologies, along with capacities to adjust internal commitments to future cooperation, agents otherwise marginalized by past mistakes or bad information are able to again contribute to cooperative endeavors. As a result, these agents learn to share in the mutual benefits of goals unassailable to the isolated individual and realizable only through more complex social interaction, confirming the precedence of prosocial capacities in the evolution of ethics (cf. Han et al. 2011, 2012, 2013; Han 2013).

Pereira et al.'s research pursues a bottom-up explanation for the emergence of morality over evolutionary time. Their work confirms that agents with native capacities to better cooperate outperform those without, both in pairwise situations and in common good settings (Han et al. 2017; Martinez-Vaquero et al. 2015, 2017). It leads Pereira et al. to conclude that evolved cognitive capacities facilitating cooperation induce the emergence of what we recognize as morality in human populations (as opposed to stable cultural practices as systems of ethics understood as rules and institutions coming first, inducing cooperation instead, cf. Pereira and Saptawijaya 2015; Han and Pereira 2018). This result has important implications for moral education in youth for example, illustrating at once how research at this high level of organization can inform research at lower levels of organization in constructive ways. Moreover, it confirms that lower-level dynamics are critical to understanding social norms and commensurate public policies.

4.2 Sun's cognitive social sciences

The upshot of cognitive modeling or computational models "in the broad sense of the term" according to Sun et al. (2005) is that these models serve as operational frameworks—"broad, generic theories of cognition" (Sun et al. 2005, p. 616)—for the structured interpretation of "a vast amount of data" generated by the cognitive and social sciences (Sun et al. 2005, p. 615). Where Pereira et al. focus on the consequences of routine action over generations of psychologically simple moral agents, Sun et al. focus on an equally vast area, the psychological processes that render such actions, rules expressing them and further their revision through the interplay of different specialized modules of information processing constitutive of more psychologically realistic learning agents. Bridging the second and third of four levels of cognitive model as iterated at the beginning of the second section of this paper, much of Sun's research focuses on multi-agent systems, social interaction, and prosocial motivation including aspects overlapping Pereira et al.'s work at higher levels of organization such as social stabilizing capacities involving intention recognition. With a resolution on cognitive mechanisms beneath the lower bounds of their evolutionary framework, however, Sun's trademark Clarion computational architecture also resolves cognition at levels of organization overlapping in part with Tani et al.'s focus on fundamental neurodynamics, for example attending to social autonomy and emotion among other psychological constructs (cf. Sun 2002, 2009, 2013, 2016, 2017, 2018a, b; Sun and Naveh 2004; Sun et al. 2016).

Consider, for example, the relationship between Sun's research program and Pereira's in greater detail. Clarion is motivated by 11 primary drives, of which many correspond to native capacities to cooperate on Pereria et al.'s model. For example, "similance"—the drive for one to identify with, and to emulate others—along with other primary drives including those to affiliate with others and to belong to groups, to avoid harmful situations, to resist control and to ensure that one's self and others are treated fairly, all work together to demonstrate a more detailed model of moral agent psychology unnecessary at Pereira et al.'s scale of evolutionary game theory (see Sun 2017, Table 1, p. 6, for a most recent summary of drives in Clarion). Clarion's psychological realism sets it apart from Pereira et al.'s model agents in other ways as well. For example, Clarion has been assessed for consciousness alongside competing architectures, and found to represent aspects of consciousness including qualia (Gok and Sayan 2012). However, like Pereira et al.'s model agents, Clarion does not aim to replicate human biological cognition or to capture the principles of its self-organization, and rather works at the level above physiological processes on Sun et al.'s four-tier scheme, at the level of psychological processes instead.

At the center of Sun's study of psychological processes is a "causal nexus" of activity between the implicit and explicit modes of information processing characteristic of hybrid systems of which Clarion is an example. Hybrid systems represent higher and lower cognitive functions in different ways, and these can interact with each other in up- and downstream processes. Clarion consists of a number of hybrid subsystems (cf. Sun 2002) whose bottom levels mediate routine action and encode regularities from which top-level rules are extracted and which are then applied top-down in the direction of future action (Sun et al. 2001; Sun 2016). As the model agent learns (bottom-up) to autonomously specify and modulate goals (which can also be learned top-down), relatively stable constructs within the cognitive architecture amount to what we recognize as "personality" in human beings (cf. Sun and Wilson 2014). Stable personalities make for predictable intentions, which Sun and Pereira capture in ways that complement one another. Sun et al.'s view complements that established by Pereira et al. by extending insight into those cognitive capacities which contribute to conventions and institutions grounding lasting social–political systems (and so mechanisms of artificial selection influencing human evolution at the same time) as these, at a finer grain of analysis, are strengthened into agent-specific personalities in a context-dependent manner. Moreover, occupying this middle space between biological constitution and cultural realization, Sun's program affords an inroad into larger systems for the influence of cognitive processes resolved by models at lower levels of organization designed to capture the dynamics of such personality formation internal to the individual agent, itself.

4.3 Tani's fundamental neurodynamics

Predictive coding along with active inference and the FEP constitute an important approach to understanding how cognition works at different levels of organization, serving as a broad framework according to which vast amounts of data from cognitive and social sciences can be interpreted. At the level of neurodynamics, its efficacy is confirmed by how well Tani et al.'s computational models demonstrating these principles articulate biological cognition in neurorobots. This section reviews Tani et al.'s program, and the next section sketches how the principles underwriting this research may be extended to cognitive models at higher levels of organization in their integration toward platforms designed to inform public policy.

Recalling the brush with dynamic systems theory and predictive coding in Sect. 3, on Tani et al.'s program a learning agent develops an "internal model" of the world as a set of self-organized dynamic attractors (Tani 1996; Tani and Nolfi 1999) toward which future actions aim. These aims are then challenged in the conflictive interaction between top-down and bottom-up processing streams (note the parallel with Sun's "causal nexus") as the perceived world deviates from model projections, resulting in an unstable "critical" state followed by the effortful return to stable coherency with the perceived reality as this internal model is recomposed (cf. Tani 2007). This process is repeated over time in various environments as the agent learns to achieve different goals, and the result is an artificial embodiment of the principles that account for similar processes in biological cognition (White and Tani 2016, 2017; Tani and White 2017).

For example, employing a relatively simple architecture using RNNs tuned to different timescales, with the lower level at a shorter timescale sensitive to rapid changes in the environment, and the higher level at a slower timescale able to extract longer-standing patterns from the same input (reflecting its predictive coding framework), Nishimoto and Tani (2009) demonstrated the development of a stable functional hierarchy whereby primitive behaviors that develop early on in the lower levels are composed into more complicated action routines in the higher level as the agent learns to achieve increasingly challenging goals during later stages, corresponding to Piaget's constructivist developmental psychology (cf. Piaget 1954). Namikawa et al. (2011) further relate these results to the developmental process of the dynamical hierarchy involving the prefrontal cortex, supplementary motor area and primary motor cortex in human beings during spontaneous composition of complex actions from primitives, as in both computational and biological systems the prefrontal areas develop similarly and—depending on the conditions of this development—deliver similar patterns of behavior.

Noteworthy is that Tani's models are not "hybrid" like Sun's, as higher and lower levels share the same metric space, i.e., they do not represent information in different ways, but rather find different patterns in different aspects of ongoing information processing in the same ways. In this way, Tani et al.'s approach to cognitive modeling is able to complement investigations undertaken on Sun et al.'s psychological approach in terms of biologically plausible—rather than psychologically plausible—dynamics. For instance, complementing Sun's (2013) account of creativity due to subsymbolic dynamics in hybrid models, Tani et al. have also investigated how actions are learned and why novel actions are composed. In a social situation, Ito and Tani (2004) employed a mirror neuron model to investigate how a complex action routine can be encoded as a single "chunk" when agent/environment dynamics are predictable, and how the resulting single seamless operation can then be resegmented into constitutive primitives through backpropagated prediction error when input proves unpredictable, with these pieces then autonomously recomposed into new patterns as the system attempts to restore up- and downstream coherency with perceived reality through novel action in response. With a robot motivated by this model to coordinate with a human subject, and with the human simultaneously attempting the same, Tani et al. learned that even small perturbations in the robot's actions could cause confusion in human subjects while the subjects were becoming accustomed to the robot's repertoire of learned action sequences (and vice versa). As a result, turn taking (with either robot or human subject leading action sequences) became prevalent during this mutual learning period, a fact that Tani and Ito interpreted as mutually initiated in response to the breakdown of higher-level intentional constructs (or "criticality") in both human and robot partners.

Murata et al. (2014) further investigated the proactive coordination of one's own actions with another's predictable action sequences as opposed to the reactive dynamics which come into play as a predicted action sequence is in error, with interacting agents each aiming to restore coherency between up- and downstream processes, echoing again Sun's focus on this nexus of activity. Murata et al. (2017) extends these results by developing a robot that integrates external with internal sources of information in the continuous performance of sensory-dependent and sensory-independent tasks, thereby operating both online (open to influence from the outside) and off-line (operating according to internal determinations) as task processing demands. Together, we can begin to see how cooperative interaction is motivated by shared neurodynamics, on this portrait as agents variably integrate self and other information online during coordinated interaction, pursuing best prediction from the top-down given bottom-up cues also recalling Pereira et al.'s intention recognition and abducible ends at a much finer grain of analysis.

In summary, Tani et al.'s research affords insight into aspects of the human condition that cannot be realistically represented at higher levels of organization, for example into the dynamic origins of agent autonomy. Tani et al.'s model agents learn action primitives during entrainment with their object environments. These learned primitives are then recomposed in response to changing conditions to align information processing streams and maintain a stable internal world model (Tani 2016, see Chap. 8, Sect. 4). The point here is twofold. For one thing, the primitives employed in novel action composition are not freely chosen, but are limited to those already learned. For another, impetus to recompose complex action routines emerges also not as a matter of choice, but rather in response to changes as the project model deviates from perceived reality. On Tani's account of these dynamics, an agent may feel as if he or she, or it, is radically free to compose novel intentions ex nihilo, as if free to do anything, but this is only due to the lack of access to the processes underlying the composition of actions (see Tani 2016, Chap. 10 for extensive discussion). Given this access, computational models of higher levels of organization may demonstrate how free action may be directed by public policy designed to optimize social conditions for maximal human creativity, cognizant of stressors which, when nearing certain thresholds, may increase rather than decrease adaptability to changing environmental conditions, perhaps by expanding the bounds of social cohesion by encouraging the development of capacities for coordination, for example.

5 Discussion

Briefly consider how a coordinated development of the three programs reviewed above into a single platform for the information of public policy might play out. Through the extension of fundamental insights from the free energy principle and predictive coding into Sun et al.'s models of social cognition, simulations should prescribe that agents entertain relationships only with those others intent on actions contributing to the minimization of uncertainty. As we extend the results of these models into Pereira et al.'s research, we should find stable arrangements potentiated, arrived at and maintained through the institutions of apology and forgiveness, as well as through promise keeping and transparency of intention. In holding social systems together, we may identify such routine cognitive agency as virtuous, and their contraries vicious. Through simulation of critical periods at this level, systematic reconciliation of the vicious with the virtuous might be recommended, and global agreement protocols permitting a systematic transition toward more stable, more cooperative, less exploitive arrangements may result without the violence that had punctuated historical transitions through similarly critical periods.

These guidelines may then be passed down through Sun et al.'s level of social agency at the level of rules, for example that agents should act to stabilize expectations within parameters conducive to cooperation and should not engage in deceit, withholding or manipulating information for selfish gain. Simulations at this level should deliver advice on how individuals might best reconcile prior understandings with that understanding necessary to motivate personal changes, facilitating social transitions to institutional arrangements through which goal conditions are realized. Selected cases may then be passed down to the level of internal dynamics as revealed through Tani et al.'s research, and here we may gain some insight into the context-dependent stress that an agent may experience during a given transition along with possible strategies for creatively turning this stress forward into constructive social contributions. With this information, people may be able to take steps to not only minimize the stress of change, but also to maximize the potential to develop healthy responses to it, responses that minimize uncertainty going forward and that at the same time maximize their own free agency, to determine for themselves how they may contribute to pro-social transitions in the interests of justice and the good life in general. These possible solutions may then be passed upward to Sun's and then to Pereira's levels of analysis, therein tested to see how they contribute to the overall stability of the systems that they aim to affect, and then to see if they should inform general principles over the generations that may be necessary to see these affects optimally realized.

From here, simulations may continue, be refined, or restarted using different parameters for evaluation of possible ends given different starting points, for example those representing possible crisis, and a science of suites of simulations such as the one sketched above may establish itself. With AI technology developed in this way, people may eventually be able to choose the world in terms of which they would like to live, and use these simulations to help to plot how to get there, openly, responsibly, at the level of the individual agent in real time and in constant view of the implications of one's actions for broader society, as well as for civilization as a whole.

Again, one upshot of this approach is that it takes advantage of increasing momentum in the integration of the cognitive with the social sciences. By beginning with ongoing research programs, the present proposal avoids some costs of development, at the same time leveraging validation of established models to facilitate acceptance of the use of such tools in the information of public policy going forward. Some further upshots of the current approach to integrate across select levels of organization are that computational costs may be lowered relative to more complex simulations intended for other purposes, and that platforms may be more rapidly developed beginning with existing research programs than by developing such predictive simulations ab initio. All together, these advantages go a long way to putting informative simulations within reach, such that, reinforcing Sun (2018b) on this point, psychologically realistic computational modeling should become a core component of future public policy education.

All of this aside, the most important upshot of the present proposal is that, at every level of analysis, there should be an accounting of energetic requirements and return into the ecology. This accounting should translate stress as understood on Tani's framework into social dynamics as represented in Sun's and Pereira's models, and we may compare the efficiencies of social arrangements operating under different types and rates of stress. In short, as stressors mount, performance becomes erratic. Some stressors even result in pathologies (as in Yamashita and Tani 2012, for example). Systems may run less efficiently or break down altogether. Policies must be adjusted, yet change must be accommodated. There are costs every step of the way. Modeling at select levels may inform policy makers on how to proactively mitigate these costs over near and long terms in such a way that implications may be made explicit, affording an opportunity for anticipatory response. For instance, impacts on the environment of a given policy initiative may be characterized at near and long terms, with more rapidly realized benefits evaluated against stress on affected populations, such that actions may be adjusted in coordination with other aspects of the larger system in terms of which the given policy is embedded. Simulations may confirm for instance that as industry slows, pollution declines, yet that standards of living eventually rise as, through a balance of high technology and traditional methods, people become wiser, healthier, live longer with greater leisure and with more time to reflect on the beauties of flourishing natural systems such as their own, and, without the chemical pollution and radiation threatening the natural world as is the case today, they flourish as well. However, moves must be made in the interim to pave the way to such a possible future. Simulations such as those subject of the present proposal may afford a survey of the landscape ahead, such that transitions through difficult passages may be most assured in getting there.

6 Conclusion

With the present proposal in mind, we may read Putin's introductory prediction another way. Instead of increased capacities for violence and coercion, the mastery of AI and robotics may present us with opportunities to stabilize rather than to destabilize relations between seemingly disparate interests. This interpretation makes sense. For one thing, it accords with what cognitive science tells us about the nature of cognition. As seen in the brief review of Friston et al.'s research, and as confirmed in Tani et al.'s neurorobots, the human brain should not optimize to the existential uncertainty that results from a "race to the precipice" with wheels greased by headline garnering intelligent machine-mediated warfare (cf. Armstrong et al. 2016). Such a condition represents a diseased state, stressed to the breaking point. Rather, healthy cognition involves the minimization of this uncertainty. In this, we may find an indication as to what should be done.

The emphasis of the preceding paper has not been on an AI arms race as competition, but rather on AI as affording avenues for peaceful cooperation toward ideal ends over the generations of human life and action that may be required to get us there. AI, to borrow from Ramstead et al. (2018), may help us to coordinate large-scale and long-term "phase-transitions" from unsustainable to sustainable, from unjust to just institutional arrangements proactively, openly, and in a non-coercive manner. Through the mastery of artificial intelligence thus, reason may yet rule the world after all.

Acknowledgements The author thanks Luis Pereira, Ron Sun, and Jun Tani for help in the preparation of this draft, as well as Karamjit Gill for his constant support. Thanks are also due to Lorenzo Magnani for advice on abduction, the anonymous reviewers of this journal and attendees of the October 18, 2018 Tech and Values colloquium at the University of Twente for insights and objections informing final revisions.

References

Aliseda A (2006) Abductive reasoning. Logical investigations into discoveryand explanation. Springer, Berlin

Armstrong S, Bostrom N, Shulman C (2016) Racing to the precipice: a model of artificial intelligence development. AI Soc 31:2:201–206

Cockshott P, Renaud K (2016) Humans, robots and values. Technol Soc 45:19–28

Fitzgerald THB, Dolan RJ, Friston K, Fitzgerald THB, Dolan RJ (2015) Dopamine, reward learning, and active inference. Front Comput Neurosci 9:1–16

Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11:127–138

Friston K (2012) A free energy principle for biological systems. Entropy 14(12):2100–2121

Friston K (2013) Life as we know it. J R Soc Interface 10:86–97

Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo ODJ G (2016) Active inference and learning. Neurosci Biobehav Rev 68:862–879

Gabbay DM, Woods JH (2005) A practical logic of cognitive systems: insight and trial. Elsevier, Amsterdam

Gok SE, Sayan E (2012) A philosophical assessment of computational models of consciousness. Cogn Syst Res 17–18:49–62

Han TA (2013) "Intention recognition, commitments and their roles in the evolution of cooperation: from artificial intelligence techniques to evolutionary game theory models" SAPERE 9. Springer, Berlin

Han TA, Pereira LM (2013) State-of-the-art of intention recognition and its use in decision making. AI Commun 26:237–246

Han TA, Pereira LM (2018) Evolutionary machine ethics. In: Bendel O (ed) Handbuch Maschinenethik. Springer reference Geisteswissenschaften. Springer, Wiesbaden, pp 1–25

Han TA, Pereira LM, Santos FC (2011) Intention recognition promotes the emergence of cooperation. Adapt Behav 3:264–279

Han TA, Pereira LM, Santos FC (2012) Corpus-based intention recognition in cooperation dilemmas. Artif Life 18(4):365–383

Han TA, Pereira LM, Santos FC, Lenaerts T (2013) Good agreements make good friends. Sci Rep 3:2695. https://www.nature.com/articles/srep02695. Accessed 18 Oct 2018

Han TA, Pereira LM, Santos FC, Lenaerts T (2015) Emergence of cooperation via intention recognition, commitment, and apology—a research summary. AI Commun 2:709–715

Han TA, Pereira LM, Lenaerts T (2017) Evolution of commitment and level of participation in public goods games. Auton Agents Multi-Agent Syst 31(3):561–583

Ito M, Tani J (2004) On-line imitative interaction with a humanoid robot using a mirror neuron model. Proc IEEE Int Conf Robot Autom 2:1071–1076

Magnani L (2017) The abductive structure of scientific creativity: an essay on the ecology of cognition. Springer, Switzerland

Marr D (1982) Vision. WH Freeman, New York

Martinez-Vaquero LA, Han TA, Pereira LM, Lenaerts T (2015) Apology and forgiveness evolve to resolve failures in cooperative agreements. Sci Rep 5:10639

Martinez-Vaquero LA, Han TA, Pereira LM, Lenaerts T (2017) When agreement-accepting free-riders are a necessary evil for the evolution of cooperation. Sci Rep. https://doi.org/10.1038/s41598-017-02625-z

Murata S, Arie H, Ogata T, Sugano S, Tani J (2014) Learning to generate proactive and reactive behavior using a dynamic neural network model with time-varying variance prediction mechanism. Adv Robot 28:1189–1203

Murata S, Masuda W, Tomioka S, Ogata T, Sugano S (2017) Mixing actual and predicted sensory states based on uncertainty estimation for flexible and robust robot behavior. In: Lintas A, Rovetta

S, Verschure P, Villa A (eds) Artificial neural networks and machine learning—ICANN 2017. ICANN 2017. Lecture notes in computer science, vol 10613. Springer, Cham, pp 11–18

Namikawa J, Nishimoto R, Tani J (2011) A neurodynamic account of spontaneous behaviour. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1002221

Newell A, Simon H (1976) Computer science as empirical inquiry: symbols and search. Commun ACM 19:113–126

Nishimoto R, Tani J (2009) Development of hierarchical structures for actions and motor imagery: a constructivist view from synthetic neuro-robotics study. Psychol Res 73:545–558

Paavola S (2006) On the origin of ideas: an abductivist approach to discovery. Department of Philosophy, University of Helsinki, Helsinki

Pereira LM, Saptawijaya A (2015) Bridging two realms of machine ethics. In: White J, Searle R (eds) Rethinking machine ethics in the age of ubiquitous technology. IGI Global, Hershey

Pereira LM, Saptawijaya A (2016) Programming machine ethics. Springer SAPERE series 26. Springer, Berlin

Pereira LM, Saptawijaya A (2017) Counterfactuals, logic programming and agent morality. In: Urbaniak R, Payette G (eds) Applications of formal philosophy: the road less travelled. Springer logic, argumentation and reasoning series. Springer, Berlin

Pereira LM, Dell'Acqua P, Pinto AM, Lopes G (2013) Inspecting and preferring abductive models. In: Nakamatsu K, Jain LC (eds) The handbook on reasoning-based intelligent systems. World Scientific Publishers, Singapore, pp 243–274

Peters A, McEwen BS, Friston K (2017) Uncertainty and stress: why it causes diseases and how it is mastered by the brain. Prog Neurobiol 156:164–188

Piaget J (1954) The construction of reality in the child. Basic Books, New York

Ramstead M, Badcock P, Friston K (2018) Answering Schrödinger's question: a free-energy formulation. Phys Life Rev. https://doi.org/10.1016/j.plrev.2017.09.001

RT (2017) 'Whoever leads in AI will rule the world': Putin to Russian children on Knowledge Day. https://www.rt.com/news/401731-ai-rule-world-putin/. Accessed 12 Oct 2018

Saptawijaya A, Pereira LM (2013) Towards practical tabled abduction in logic programs. In: Correia L, Reis LP, Cascalho J (eds) Progress in artificial intelligence. EPIA 2013. Lecture notes in computer science, vol 8154. Springer, Berlin, pp 223–234

Saptawijaya A, Pereira LM (2018) From logic programming to machine ethics. In: Bendel O (ed) Handbuch Maschinenethik. Springer, Berlin

Schwartenbeck P, FitzGerald THB, Mathys C, Dolan R, Kronbichler M, Friston K (2015) Evidence for surprise minimization over value maximization in choice behavior. Sci Rep 5:1–14

Shipp S, Adams RA, Friston KJ (2013) Reflections on agranular architecture: predictive coding in the motor cortex. Trends Neurosci 36(12):706–716

Simon HA (1977) Models of discovery: and other topics in the methods of science. D Reidel Pub Co, Dordrecht

Sun R (2002) Duality of the mind. Lawrence Erlbaum Associates, Mahwah

Sun R (2009) Motivational representations within a computational cognitive architecture. Cogn Comput 1(1):91–103

Sun R (2012) Grounding social sciences in cognitive sciences. MIT Press, Cambridge

Sun R (2013) Autonomous generation of symbolic representations through subsymbolic activities. Philos Psychol 26(6):888–912. https://doi.org/10.1080/09515089.2012.711035

Sun R (2016) Anatomy of the mind: exploring psychological mechanisms and processes with the Clarion cognitive architecture. Oxford University Press, New York

Sun R (2017) Potential of full human–machine symbiosis through truly intelligent cognitive systems. AI Soc. https://doi.org/10.1007/s00146-017-0775-7

Sun R (2018a) Intrinsic motivation for truly autonomous agents. In: Abbass H, Scholz J, Reid D (eds) Foundations of trusted autonomy. Springer, Berlin, pp 273–292

Sun R (2018b) Cognitive social simulation for policy making. Policy Insights Behav Brain Sci 5(2):240–246

Sun R, Naveh I (2004) Simulating organizational decision-making using a cognitively realistic agent model. J Artif Soc Soc Simul. http://jasss.soc.surrey.ac.uk/7/3/5.html. Accessed 15 Dec 2017

Sun R, Wilson N (2014) A model of personality should be a cognitive architecture itself. Cogn Syst Res 29:1–30

Sun R, Merrill E, Peterson T (2001) From implicit skills to explicit knowledge: a bottom–up model of skill learning. Cogn Sci 25(2):203–244

Sun R, Coward LA, Zenzen MJ (2005) On levels of cognitive modeling. Philos Psychol 18(5):613–637

Sun R, Wilson N, Lynch M (2016) Emotion: a unified mechanistic interpretation from a cognitive architecture. Cogn Comput 8:1:1–14

Tani J (1996) Model-based learning for mobile robot navigation from the dynamical systems perspective. IEEE Trans Syst Man Cybern Part B-Cybern 26(3):421–436

Tani J (2007) On the interactions between top-down anticipation and bottom-up regression. Front Neurorobot 1:1–10

Tani J (2016) Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena. Oxford University Press, New York

Tani J, Nolfi S (1999) Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems. In: Pfeifer R, Blumberg B, Meyer JA, Wilson SW (eds) In: Proceedings of 5th international conference on simulation of adaptive behavior. MIT Press, Massachusetts, pp 270–279

Tani J, White J (2017) From biological to synthetic neurorobotics approaches to understanding the structure essential to consciousness, part 2. APA Newsl Philos Comput 16(2):29–41

White J (2016) Simulation, self-extinction, and philosophy in the service of human civilization. AI Soc 31(2):171–190

White J, Tani J (2016) From biological to synthetic neurorobotics approaches to understanding the structure essential to consciousness, part 1. APA Newsl Philos Comput 16(1):13–23

White J, Tani J (2017) From biological to synthetic neurorobotics approaches to understanding the structure essential to consciousness, part 3. APA Newsl Philos Comput 17(1):11–22

Yamashita Y, Tani J (2012) Spontaneous prediction error generation in Schizophrenia. PLoS One 5(5):e37843. https://doi.org/10.1371/journal.pone.0037843